2016

Selected topics in measurement error and functional data analysis

Yuhang Xu
Iowa State University

Selected topics in measurement error and functional data analysis

by

Yuhang Xu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Yehua Li, Major Professor

Songxi Chen

Jae-kwang Kim

William Q. Meeker

Dan Nettleton

Iowa State University

Ames, Iowa

2016

TABLE OF CONTENTS

ABSTRACT

Measurement error frequently occurs in scientific studies when precise measurements of variables are unavailable or too expensive. It is well-known that in regression models ignoring measurement error leads to biased estimation of regression coefficients. In this thesis, we propose semiparametric methods to correct the bias and improve the efficiency of estimation under two frameworks. Firstly, for proportional hazards models with measurement error in covariates, we propose a new class of semiparametric estimators by solving estimating equations based on the semiparametric efficient scores. The baseline hazard function, the hazard function for the censoring time, and the distribution of the true covariates are all treated as unknown infinite dimensional. The proposed estimators prove to be locally efficient. Secondly, for a general regression model when error-prone surrogates of true predictors are collected in the primary data set while accurate measurements of the predictors are available only in a small validation data set, we propose a new class of semiparametric estimators for the regression coefficients based on expected estimating equations. The measurement error model is calibrated nonparametrically using a kernel smoothing method. We prove that the proposed estimators are consistent, asymptotically unbiased and normal in both scenarios.

Functional data appear more and more often in scientific fields due to technological advances. In functional data analysis (FDA), function principal components analysis (FPCA) has become one of the most important modeling and dimension reduction tools. Motivated by a recent root image study in plant science where the data have a natural three-level nested hierarchical structure, we analyze the data using multilevel FPCA. We estimate the covariance function of the functional random effects by a fast penalized

tensor product spline approach, perform multilevel FPCA using the best linear unbiased predictor of the principal component scores, and improve the estimation efficiency by an iterative algorithm. We choose the number of principal components using a conditional Akaike Information Criterion and test the effect in the mean function using a generalized likelihood ratio test statistic based on the marginal likelihood and the conditional likelihood. Extensive simulation studies have been carried out to evaluate the validity of our proposed methods.

CHAPTER 1. OVERVIEW

1.1 Measurement error

A common problem in statistics is to make inference about the relationship between a response variable and predictors. In many scientific studies, however, surrogates of the predictors are collected because precise measurements of the true predictors are either unavailable or too expensive. For example, in Chapter 2, Xu, Li, and Song (2016) analyzed a data set collected in HIV clinical trials where CD4 cell counts cannot be measured accurately and are hence subject to measurement errors; in our motivating data in Chapter 3, one of the key predictors, the body mass index (BMI), is calculated based on self-reported weight and height and is subject to measurement errors. Measurement error in covariates has several bad effects such as causing bias in parameter estimation and leading to a loss of power for detecting the relationship between the response and covariates.

To perform a measurement error analysis, usually extra information except the response and the surrogates of predictors is needed. Usually, there are two sources of data which allow us to correct the effects mentioned above. The first type of data is called replication data where replicates of the surrogates are available for each subject. In the data analysis conducted in Chapter 2, CD4 cell counts are measured multiple times for each patient, so replication data are available. Usually, an assumption about the surrogates and the true predictors needs to be made. For example, the most standard one is the classic measurement error model (Fuller, 1987; Carroll et al., 2006) which assumes

that the surrogates equal to the sum of the true predictors and measurement errors. The second type of data is called validation data where true predictors are available in an extra validation data set. For example, in Chapter 3, we analyze a data set from the Korean Longitudinal Study of Aging (KLoSA) where precise physical measurements on the height and weight are available for a part of the subjects. For this type of data, as we stated in Chapter 3, the traditional measurement error model assumption can be relaxed and a nonparametric measurement error model can be considered using the validation data.

In survival analysis, the proportional hazards model (Cox, 1972) is the most widely used model. When all or a subset of the covariates are measured with error, lots of methods have been proposed to deal with measurement errors in the proportional hazards model. For example, regression calibration (Prentice, 1982; Wang et al., 1997), SIMEX (Li and Lin, 2003), likelihood-based methods (Hu, Tsiatis, and Davidian, 1998; Su and Wang, 2012), Bayesian methods (Cheng and Crainiceanu, 2009) and score estimating equation methods. However, none of those methods leads to semiparametric efficient estimation of the regression coefficients in the proportional hazards model. Based on semiparametric efficient scores, in Chapter 2, we partially solve this problem by proposing locally efficient semiparametric estimators for the regression coefficients. Our methodology is partly motivated by Tsiatis and Ma (2004) which proposed locally efficient semiparametric estimators in the general setting of functional measurement error models.

When validation data are available, for a general regression problem which assumes that the conditional density of the response given the predictors follows a known parametric form, Chapter 3 considers semiparametric estimation of the regression parameters. Although there are some literature on semiparametric methods based on the likelihood or score equations and they consider a nonparametric measurement error model calibrated using the validation data, they all have some limitations. For example, Carroll

and Wand (1991) and Wang and Wang (1997) limit their focus to logistic regression models; Pepe and Fleming (1991) requires the surrogates to be categorical; the method proposed in Wang and Yu (2007) is inconsistent in general. Based on expected estimating equations (Wang and Pepe, 2000), we propose kernel-based semiparametric estimators for the regression coefficients to overcome all the limitations mentioned above.

1.2 Functional data analysis

In many scientific fields, such as biometrics, econometrics, plant science, etc., data are often collected over a continuum of time or space. For example, in a root image study as we described in Chapter 4, plant scientists at the University of Wisconsin-Madison used digital cameras to measure root tips angles of each maize seed every 3 minutes for a total duration of 3 hours. This type of data are named as functional data because it is often assumed that they represent a sample of i.i.d. smooth random functions over a specific time or space. Partly due to the increasing availability of functional data, functional data analysis, often called FDA, has received considerable attention in the last decade and has become one of the most active areas in modern statistics. Details about the methodology development in FDA can be found in Ramsay and Silverman (2005).

In FDA, functional principal component analysis (FPCA) proves to be one of the most important modeling and dimension reduction tools. In FPCA, based on the Karhunen-Loève expansion and the selection of the number of principal components, random trajectories are usually reduced to a linear combination of few functional principal component scores and eigenfunctions which explain major amount of variation of the random trajectories. Among the methods proposed in FPCA, the method proposed by Yao, Müller, and Wang (2005), principal components analysis through conditional expectation (PACE), has become one of the most popular methods. Some theoretical properties of FPCA using PACE are investigated by Hall, Müller, and Wang (2006) and Li and Hsing (2010).

There have been some recent research in hierarchical or multilevel functional data analysis. For example, Di et al. (2009) studies two-level hierarchical functional data from a sleep heart health study and Li, et al. (2015) analyzes three-level functional data from an exercise intervention trial. The data motivate our study in Chapter 4 also have a three-level hierarchical structure. In the current literature on hierarchical FPCA, including Di et al. (2009) and Li, et al. (2015), they select the number of components subjectively using an ad hoc "percentage of variation explained" (PVE) method which is very subjective and often leads to a wrong model as we described in Chapter 4. For this reason, a data-driven method for the model selection in multilevel FPCA is more than necessary. In Chapter 4, we extend the recent work of Li, Wang, and Carroll (2013) for independent functional data to the hierarchical setting and propose a data-driven method based on a conditional Aikaike information criterion (AIC). There is relatively little work on nonparametric inference for hierarchical functional data, but in order to test the moon phase effect on the maize seeds gravitropism, it is interesting to compare a full model where the mean function is bivariate and depends on both the measuring time and the lunar day and a reduced model where the mean function only depends on the measuring time. Motivated by this scientific question, we propose generalized likelihood ratio (GLR) test statistics (Fan, Zhang, and Zhang, 2001) for this type of test in hierarchical functional data in Chapter 4. The GLR test statistics are based on the marginal likelihood, conditional likelihood and working independence, respectively.

CHAPTER 2. LOCALLY EFFICIENT SEMIPARAMETRIC ESTIMATORS FOR PROPORTIONAL HAZARDS MODELS WITH MEASUREMENT ERROR

A paper published in *Scandinavian Journal of Statistics*

Yuhang Xu[1], Yehua Li[2], and Xiao Song[3]

Abstract

We propose a new class of semiparametric estimators for proportional hazards models in the presence of measurement error in the covariates, where the baseline hazard function, the hazard function for the censoring time, and the distribution of the true covariates are considered as unknown infinite dimensional parameters. We estimate the model components by solving estimating equations based on the semiparametric efficient scores under a sequence of restricted models where the logarithm of the hazard functions are approximated by reduced rank regression splines. The proposed estimators are locally efficient in the sense that the estimators are semiparametrically efficient if the distribution of the error-prone covariates is specified correctly, and are still consistent and asymptotically normal if the distribution is misspecified. Our simulation studies show that the proposed estimators have smaller biases and variances than competing methods. We further illustrate the new method with a real application in an HIV clinical trial.

[1]Primary researcher and author, Graduate student, Department of Statistics, Iowa State University.
[2]Author for correspondence, Associate Professor, Department of Statistics, Iowa State University.
[3]Associate Professor, Department of Epidemiology and Biostatistics, University of Georgia.

2.1 Introduction

The proportional hazards model or the Cox model (Cox, 1972) is the most widely used model in survival analysis. When all or a subset of the covariates are measured with error, it is well known that the naive method which substitutes the mismeasured values for the true covariates leads to biased estimation (Prentice, 1982). There are a large amount of methods in the literature on measurement error problems for the Cox model, including approximation methods such as regression calibration (Prentice, 1982; Wang et al., 1997) and SIMEX (Li and Lin, 2003), likelihood-based methods (Hu, Tsiatis, and Davidian, 1998; Su and Wang, 2012), Bayesian methods (Cheng and Crainiceanu, 2009) and score estimating equation methods. Typical score methods include the conditional score method (Tsiatis and Davidian, 2001; Song, Davidian, and Tsiatis, 2002) and the corrected score method (Nakamura, 1992; Huang and Wang, 2000; Song and Huang, 2005). The readers are referred to Carroll et al. (2006) for a comprehensive account of these methods.

The method we propose in this paper is most closely related to score estimating equation methods. Unlike likelihood-based methods, score methods do not assume a parametric model for the distribution of the true covariate. However there are still some drawbacks for existing score methods. For example, the conditional score approach relies on the existence of a complete and sufficient statistic and therefore is not applicable to situations where such a statistic cannot be easily obtained. One simple example where a complete and sufficient statistic does not exist is the quadratic logistic regression model described in Tsiatis and Ma (2004). More importantly, semiparametric efficiency has not yet been established for any existing score estimating equation methods in survival analysis.

Tsiatis and Ma (2004) proposed locally efficient semiparametric estimators in the general setting of functional measurement error models. Their method is based on semi-

parametric efficient scores rather than complete and sufficient statistics and therefore is more general. Their method was first proposed for parametric regression models, where the response depends on a parametric function of the true covariates, and was further extended by Ma and Carroll (2006) to a class of semiparametric regression models, where the response variable Y depends on a parametric form of an error-prone covariate $\mathbf{X}$ and a nonparametric function of an error-free univariate covariate $\mathbf{Z}$. Ma and Carroll (2006) proposed a backfitting algorithm, where the nonparametric function was estimated by a local estimating equation. In particular, they applied their method to the generalized partially linear models as a special case.

Unlike the models considered in Ma and Carroll (2006), the likelihood of the Cox model depends on an unknown baseline hazard function, the hazard function for the censoring time, and their integrals, i.e., the corresponding cumulative hazard functions. Constructing local estimating equations for the Cox model is difficult especially when some of the covariates are measured with error. Our strategy is to apply the method of Tsiatis and Ma (2004) to a restricted model where the logarithms of the baseline hazard function and the censoring hazard function are both approximated by regression splines (Zhou, Shen, and Wolfe, 1998; Zhu, Fung, and He, 2008). By slowly increasing the rank of the spline approximation with the sample size, the efficient score under the restricted model becomes an estimating equation with diverging number of covariates (Wang, 2011). There has also been a large amount of literature on estimating the baseline hazard function using splines, including Kooperberg, Stone, and Truong (1995), and Huang (1996). However, none of them has dealt with efficient estimation in measurement error problems.

The rest of the paper is organized as follows. We introduce the model framework in Section 2.2.1 and propose a class of locally efficient estimators under a class of broadly defined sub-models in Section 2.2.2. We discuss the asymptotic properties of the resulting estimators while letting the rank of the spline approximation increase with the sample

size in Section 2.2.3 and provide a specific implementation strategy based on spline approximations in Section 2.2.4. We provide two simulation studies in Section 2.3 and a real data application in AIDS clinical trials in Section 2.4 to illustrate the proposed method. Some concluding remarks are provided in Section 2.5, and all technical proofs are presented in the Appendix, Section 2.7.

2.2 Locally efficient semiparametric estimation

2.2.1 Model specification and assumptions

We consider right censored failure time data collected from n independent subjects. For subject i, let T_i and C_i be the failure and censoring times, $\mathbf{Z}_i$ be a p_1-dimensional error-free covariate and $\mathbf{X}_i$ be a p_2-dimensional error-prone covariate. Let $V_i = \min(T_i, C_i)$ be the observed event time and $\Delta_i = I(T_i \leq C_i)$ be the failure indicator. We assume the censoring time is independent of the failure time and the covariates. However, this assumption can be relaxed (see Section 2.5). The hazard of the failure time T_i is related to $\mathbf{X}_i$ and $\mathbf{Z}_i$ through a proportional hazards regression model

$$\begin{aligned} \lambda_i(t) &= \lim_{\mathrm{d}t \to 0} \mathrm{pr}(t \leq T_i < t + \mathrm{d}t | T_i \geq t, \mathbf{X}_i, \mathbf{Z}_i)/\mathrm{d}t \\ &= \lambda(t) \exp\{g(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}_1)\}, \end{aligned} \tag{2.1}$$

where $\lambda(\cdot)$ is an unspecified baseline hazard function and $g(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}_1)$ is a known function of $\mathbf{X}_i$ and $\mathbf{Z}_i$ up to a p-dimensional parameter $\boldsymbol{\theta}_1$. The most commonly used parametric structure for $g(\cdot)$ is the linear structure, i.e. $g(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}_1) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\alpha}$ with $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$ and $p = p_1 + p_2$. However, we also allow more general parametric structures for $g(\cdot)$, such as a quadratic regression model $g(X_i, \mathbf{Z}_i, \boldsymbol{\theta}_1) = X_i\beta_1 + X_i^2\beta_2 + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\alpha}$ with $\boldsymbol{\theta}_1 = (\beta_1, \beta_2, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$ and $p = p_1 + p_2 + 1$. Instead of observing $\mathbf{X}_i$, we only observe $\mathbf{W}_i$, a surrogate for $\mathbf{X}_i$. We assume that (V, Δ) and $\mathbf{W}$ are conditionally independent given the covariates $(\mathbf{X}, \mathbf{Z})$ and the conditional density of $\mathbf{W}$ given $\mathbf{X}$ and $\mathbf{Z}$, $p(\mathbf{w} \mid \boldsymbol{x}, \mathbf{z})$, is known.

For example, under a Gaussian classical measurement error model $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U} \sim \text{Normal}(\mathbf{0}, \mathbf{\Sigma_U})$ is a measurement error independent of $\mathbf{X}$, the assumption requires the covariance matrix $\mathbf{\Sigma_U}$ to be known. This assumption can be later relaxed to allow for unknown parameters in $p(\mathbf{w} \mid \boldsymbol{x}, \mathbf{z})$, see remark 3 in Section 2.2.2.

Denote the observed data as $\mathbf{O} = (V, \Delta, \mathbf{W}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$ and the data comprising the true covariate as $\mathbf{D} = (V, \Delta, \mathbf{W}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$. Following (2.1), the probability density of $\mathbf{D}$ is

$$\begin{aligned} p(\mathbf{d}) &= [\lambda(v)\exp\{g(\boldsymbol{x}, \mathbf{z}, \boldsymbol{\theta}_1)\}]^{\delta} \exp[-\Lambda(v)\exp\{g(\boldsymbol{x}, \mathbf{z}, \boldsymbol{\theta}_1)\}] \\ &\quad \times \{\lambda_{\mathrm{C}}(v)\}^{1-\delta} \exp\{-\Lambda_{\mathrm{C}}(v)\} p(\mathbf{w} \mid \boldsymbol{x}, \mathbf{z}) p(\boldsymbol{x} \mid \mathbf{z}) p(\mathbf{z}), \end{aligned} \tag{2.2}$$

where $\Lambda(t) = \int_0^t \lambda(u)du$ is the baseline cumulative hazard function, and $\lambda_{\mathrm{C}}(\cdot)$ and $\Lambda_{\mathrm{C}}(\cdot)$ are the hazard and cumulative hazard functions for the censoring time. We collect the four infinite dimensional nuisance parameters in the model into $\boldsymbol{\eta} = \{p(\boldsymbol{x} \mid \mathbf{z}), p(\mathbf{z}), \lambda(\cdot), \lambda_{\mathrm{C}}(\cdot)\}$.

The model specified in (2.2) is a semiparametric model Tsiatis (2006) with a parametric component $\boldsymbol{\theta}_1$ and a nonparametric component $\boldsymbol{\eta}$. Define the Hilbert space H_{D} of all L_2 measurable functions $h(\mathbf{D})$ of $\mathbf{D}$ with finite variance, and equip H_{D} with the covariance inner product, $\langle h_1(\mathbf{D}), h_2(\mathbf{D})\rangle = E\{h_1(\mathbf{D})h_2(\mathbf{D})\}$ for any $h_1(\mathbf{D}),\ h_2(\mathbf{D}) \in H_{\mathrm{D}}$. The nuisance tangent space, denoted as Λ_F, is the linear subspace in H_{D} spanned by the nuisance scores of all parametric sub-models (Tsiatis and Ma, 2004). Two elements $h_1,\ h_2 \in H_{\mathrm{D}}$ are orthogonal to each other, or $h_1 \perp h_2$, if $\langle h_1, h_2\rangle = 0$. For any $h \in H_{\mathrm{D}}$ and any subspace $A \subset H_{\mathrm{D}}$, define the projection of h on A, denoted by $\Pi(h|A)$, to be the unique element $h_a \in A$ such that $h - h_a$ is orthogonal to all elements in A, or $(h - h_a) \perp A$. If $\boldsymbol{h} = (h_1, \ldots, h_r)$ is a vector of random elements in H_{D}, define $\Pi(\boldsymbol{h}|A)$ to be the vector resulted from projecting each entry of $\boldsymbol{h}$ on A. Also define the orthogonal complement of a subspace A as $A^{\perp} = \{g \in H_{\mathrm{D}} :\ g \perp A\}$.

In our problem, the nuisance tangent space can be further decomposed into four subspaces associated with the four components in $\boldsymbol{\eta}$. Let $N_{\mathrm{C}}(u) = I(V \le u, \Delta = 0)$ be the

censoring indicator by time u, $N(u) = I(V \leq u, \Delta = 1)$ be the observed failure indicator, and $Y(u) = I(V \geq u)$ be the at-risk indicator. Define martingale increments $\mathrm{d}M_{\mathrm{C}}(u) = \mathrm{d}N_{\mathrm{C}}(u) - \lambda_{\mathrm{C}}(u)Y(u)\mathrm{d}u$ and $\mathrm{d}M(u, \mathbf{X}, \mathbf{Z}) = \mathrm{d}N(u) - \lambda(u)\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\alpha})Y(u)\mathrm{d}u$. As derived in Tsiatis and Ma (2004), the nuisance tangent subspaces associated with $p(\boldsymbol{x} \mid \mathbf{z})$ and $p(\mathbf{z})$ are

$$\begin{aligned} \Lambda_{1\mathrm{D}} &= [h(\mathbf{X}, \mathbf{Z}) \in H_{\mathrm{D}} : \ E\{h(\mathbf{X}, \mathbf{Z}) \mid \mathbf{Z}\} = 0], \\ \Lambda_{2\mathrm{D}} &= [h(\mathbf{Z}) \in H_{\mathrm{D}} : \ E\{h(\mathbf{Z})\} = 0]. \end{aligned}$$

Tsiatis (2006) shows that the nuisance tangent subspaces associated with $\lambda(\cdot)$ and $\lambda_{\mathrm{C}}(\cdot)$ are

$$\begin{aligned} \Lambda_{3\mathrm{D}} &= \{\textstyle\int a(u)\ \mathrm{d}M(u, \mathbf{X}, \mathbf{Z})\}, \\ \Lambda_{4\mathrm{D}} &= \{\textstyle\int a(u)\ \mathrm{d}M_{\mathrm{C}}(u)\}, \end{aligned}$$

where $a(u)$ is any integrable function of u. It is easy to verify that the four nuisance tangent subspaces above are orthogonal to each other under the covariance inner product, and hence Λ_{D} is a direct sum of the four subspaces, i.e. $\Lambda_{\mathrm{D}} = \Lambda_{1\mathrm{D}} \oplus \Lambda_{2\mathrm{D}} \oplus \Lambda_{3\mathrm{D}} \oplus \Lambda_{4\mathrm{D}}$.

Similarly, we define the Hilbert space H of functions on the observed data. Using arguments similar to those in Tsiatis and Ma (2004), we can see that $H = E(H_{\mathrm{D}}|\mathbf{O}) = \{E(h|\mathbf{O}) : \ h \in H_{\mathrm{D}}\}$ and the nuisance tangent space based on the observed data is

$$\Lambda = E(\Lambda_{\mathrm{D}} \mid \mathbf{O}) = \Lambda_1 + \Lambda_2 + \Lambda_3 + \Lambda_4,$$

where $\Lambda_i = E(\Lambda_{i\mathrm{D}} \mid \mathbf{O})$, $i = 1, \ldots, 4$. Since $\Lambda_{2\mathrm{D}}$ and $\Lambda_{4\mathrm{D}}$ only consist of functions of the observed data, it is easy to see that $\Lambda_2 = \Lambda_{2\mathrm{D}}$ and $\Lambda_4 = \Lambda_{4\mathrm{D}}$, and they are both orthogonal to Λ_1 and Λ_3. However, Λ_1 and Λ_3 are no longer orthogonal to each other.

The observed-data score is a p-dimensional vector $E[\partial/\partial\boldsymbol{\theta}_1\{\log p(\mathbf{d})\} \mid \mathbf{O}]$. The semiparametric efficient score is the projection of the observed-data score onto $\Lambda^{\perp}$, the orthogonal complement of the nuisance tangent space (Tsiatis and Ma, 2004). In theory,

the root of the efficient score is the most efficient estimator for $\boldsymbol{\theta}_1$. However, except for a few special cases, the semiparametric efficient score usually involves the unknown nuisance parameter $\boldsymbol{\eta}$ and does not have a close form.

2.2.2 Locally efficient estimators under a restricted sub-model

To motivate our locally efficient semiparametric estimator, we consider a sub-model of (2.2) where both $\nu(t;\boldsymbol{\gamma}_1) = \log\{\lambda(t)\}$ and $\nu_{\mathrm{C}}(t;\boldsymbol{\gamma}_2) = \log\{\lambda_{\mathrm{C}}(t)\}$ are modeled parametrically. Suppose the parameters $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are of dimensions K_1 and K_2 respectively, and we will refer to this sub-model as the restricted model.

Define $\boldsymbol{\theta}_2 = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$, $K = K_1 + K_2$ and $q = p + K$. Then the restricted model is a semiparametric sub-model with a q-dimensional parametric component $\boldsymbol{\theta}$ and a nonparametric component $\boldsymbol{\eta}^{\mathrm{R}} = \{p(\boldsymbol{x} \mid \mathbf{z}), p(\mathbf{z})\}$. Under this restricted model, the nuisance tangent spaces for $\mathbf{D}$ and $\mathbf{O}$, denoted by $\Lambda_{\mathrm{D}}^{\mathrm{R}}$ and Λ^{R}, have simpler structures

$$\Lambda_{\mathrm{D}}^{\mathrm{R}} = \Lambda_{1\mathrm{D}} \oplus \Lambda_{2\mathrm{D}}, \quad \Lambda^{\mathrm{R}} = \Lambda_1 \oplus \Lambda_2,$$

where $\Lambda_{i\mathrm{D}}^{\mathrm{R}}$ and Λ_i^{R}, $i = 1, 2$, are defined as in Section 2.2.1.

In locally efficient semiparametric estimation, the underlying conditional density $p(\boldsymbol{x} \mid \mathbf{z})$ is allowed to be misspecified. Denote the possibly incorrect conditional density by $p^*(\boldsymbol{x} \mid \mathbf{z})$, expectations or conditional expectations calculated under the misspecified distribution by $E^*(\cdot)$, and expectations calculated under the true distribution by $E(\cdot)$. Define the Hilbert spaces, H_{D}^* and H^*, for $\mathbf{D}$ and $\mathbf{O}$ under the misspecification, where the inner product is defined on $E^*(\cdot)$. Similarly, $\Lambda_{\mathrm{D}}^{\mathrm{R}*} = \Lambda_{1\mathrm{D}}^* \oplus \Lambda_{2\mathrm{D}}^*$ and $\Lambda^{\mathrm{R}*} = \Lambda_1^* \oplus \Lambda_2^*$ are the nuisance tangent spaces of the restricted model under $p^*(\boldsymbol{x} \mid \mathbf{z})$. Denote Π^* as the projection operator under $E^*(\cdot)$.

The score vector based on $\mathbf{D}$ under the restricted model is

$$\mathbf{S}_{\mathrm{D},\boldsymbol{\theta}}(V, \Delta, \mathbf{X}, \mathbf{Z}) = \partial/\partial\boldsymbol{\theta}\{\log p(V, \Delta \mid \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})\},$$

and the observed-data score vector under $p^*(\boldsymbol{x} \mid \mathbf{z})$ is

$$\begin{aligned} \mathbf{S}^*_{\boldsymbol{\theta}}(\mathbf{O}) &= E^*\{\mathbf{S}_{\text{D},\boldsymbol{\theta}}(V, \Delta, \mathbf{X}, \mathbf{Z}) \mid \mathbf{O}\} \\ &= \frac{\int \mathbf{S}_{\text{D},\boldsymbol{\theta}}(V, \Delta, \boldsymbol{x}, \mathbf{Z}) p(V, \Delta \mid \boldsymbol{x}, \mathbf{Z}; \boldsymbol{\theta}_0) p(\mathbf{W} \mid \boldsymbol{x}, \mathbf{Z}) p^*(\boldsymbol{x} \mid \mathbf{Z}) \, \mathrm{d}\mu(\boldsymbol{x})}{\int p(V, \Delta \mid \boldsymbol{x}, \mathbf{Z}; \boldsymbol{\theta}_0) p(\mathbf{W} \mid \boldsymbol{x}, \mathbf{Z}) p^*(\boldsymbol{x} \mid \mathbf{Z}) \, \mathrm{d}\mu(\boldsymbol{x})}, \end{aligned} \tag{2.3}$$

where $\mathrm{d}\mu(\cdot)$ denotes the dominating measure in the domain of $\mathbf{X}$. Based on the observed score (2.3), the efficient score for the restricted model under $p^*(\boldsymbol{x} \mid \mathbf{Z})$ is

$$\mathbf{S}^*_{\text{eff}}(\mathbf{O}, \boldsymbol{\theta}) = \mathbf{S}^*_{\boldsymbol{\theta}}(\mathbf{O}) - \Pi^*\{\mathbf{S}^*_{\boldsymbol{\theta}}(\mathbf{O}) | \Lambda^{\text{R}*}\}. \tag{2.4}$$

The following Lemma follows directly from Theorem 1 of Tsiatis and Ma (2004) and provides clues to find $\Pi^*\{\mathbf{S}^*_{\boldsymbol{\theta}}(\mathbf{O}) | \Lambda^{\text{R}*}\}$.

Lemma 1 *Suppose the restricted model is true, and $p^*(\boldsymbol{x} \mid \mathbf{z})$ and $p(\boldsymbol{x} \mid \mathbf{z})$ have the same support, then almost surely, the space orthogonal to $\Lambda_1^{\text{R}*} \oplus \Lambda_2^{\text{R}*}$ is*

$$[h(\mathbf{O}) : E\{h(\mathbf{O}) \mid \mathbf{X}, \mathbf{Z}\} = \mathbf{0}].$$

By the definition of $\Lambda^{\text{R}*}$, $\Pi^*\{\mathbf{S}^*_{\boldsymbol{\theta}}(\mathbf{O}) \mid \Lambda^{\text{R}*}\}$ is of the form $E^*\{h(\mathbf{X}, \mathbf{Z}) \mid \mathbf{O}\}$. By Lemma 1, almost surely, $h(\mathbf{X}, \mathbf{Z})$ needs to satisfy the integral equation

$$E^*[\mathbf{S}^*_{\boldsymbol{\theta}}(\mathbf{O}) - E^*\{h(\mathbf{X}, \mathbf{Z}) \mid \mathbf{O}\} \mid \mathbf{X}, \mathbf{Z}] = \mathbf{0}. \tag{2.5}$$

Details of how to solve the integral equation above can be found in Section 2.2.4.

Remark 1 Under the restricted model, the efficient score in (2.4) becomes

$$\mathbf{S}^*_{\text{eff}}(\mathbf{O}, \boldsymbol{\theta}) = \mathbf{S}^*_{\boldsymbol{\theta}}(\mathbf{O}) - E^*\{h(\mathbf{X}, \mathbf{Z}) \mid \mathbf{O}\}, \tag{2.6}$$

where $h(\mathbf{X}, \mathbf{Z})$ is the solution of (2.5). Lemma 1 implies that $E\{\mathbf{S}^*_{\text{eff}}(\mathbf{O}, \boldsymbol{\theta}) \mid \mathbf{X}, \mathbf{Z}\} = \mathbf{0}$, which in turn implies $E\{\mathbf{S}^*_{\text{eff}}(\mathbf{O}, \boldsymbol{\theta})\} = \mathbf{0}$ and $\mathbf{S}^*_{\text{eff}}(\mathbf{O}, \boldsymbol{\theta})$ as an element in H is orthogonal to $\Lambda_1 + \Lambda_2$. In other words, even if $p^*(\boldsymbol{x} \mid \mathbf{Z})$ is misspecified, the estimating equation based on the score function in (2.6) is still unbiased.

Based on the efficient score (2.6), we construct the estimating equation

$$\mathbf{S}_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\mathbf{S}^*_{\text{eff}}(\mathbf{O}_i, \boldsymbol{\theta}) = \mathbf{0}, \tag{2.7}$$

where $\mathbf{O}_i = (V_i, \Delta_i, \mathbf{W}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$. Denote the solution of (2.7) by $\widehat{\boldsymbol{\theta}}_n$, which yields a partition $(\widehat{\boldsymbol{\theta}}_{1n}^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_{1n}^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_{2n}^{\mathrm{T}})^{\mathrm{T}}$ the same way as for $\boldsymbol{\theta}$. Then $\widehat{\boldsymbol{\theta}}_{1n}$ contains the estimators of the Cox regression coefficients that are of main interests.

Remark 2 The possibly misspecified conditional density $p^*(\boldsymbol{x} \mid \mathbf{z})$ is often assumed to have a parametric form $p^*(\boldsymbol{x} \mid \mathbf{z}; \boldsymbol{\tau})$, where $\boldsymbol{\tau}$ is an unknown finite-dimensional parameter of which a root-n consistent estimator $\widehat{\boldsymbol{\tau}}$ exists. Tsiatis and Ma (2004) shows that replacing $p^*(\boldsymbol{x} \mid \mathbf{z}; \boldsymbol{\tau})$ with $p^*(\boldsymbol{x} \mid \mathbf{z}; \widehat{\boldsymbol{\tau}})$ does not cause any additional variation in the final estimator $\widehat{\boldsymbol{\theta}}_n$.

Remark 3 So far, we have assumed $p(\mathbf{w} \mid \boldsymbol{x}, \mathbf{z})$ to be known. In reality, this distribution usually depends on an unknown measurement error model parameter $\boldsymbol{\gamma}_{mem}$. For example, under the classical measurement error model $\mathbf{W} = \mathbf{X} + \mathbf{U}$, $\boldsymbol{\gamma}_{mem} = \boldsymbol{\Sigma}_{\mathbf{U}}$ is the covariance matrix of the measurement error. One can estimate this parameter through replicates of $\mathbf{W}$ and plug the estimate into the estimating equation (2.7). Replicates of the surrogate exist in many applications, including the AIDS clinical trial data in Section 2.4. However, as pointed out in Ma and Carroll (2006), though this plug-in method is robust, it may cause loss of efficiency. A more efficient method is to construct an additional estimating equation for $\boldsymbol{\gamma}_{mem}$ and solve this equation with (2.7) simultaneously. See Section 3.6 of Ma and Carroll (2006) for the construction of this additional equation.

2.2.3 Locally efficient estimators based on regression splines

We now provide a more specific strategy to parameterize $\nu(t)$ and $\nu_{\mathrm{C}}(t)$, where we will use regression splines (Zhou, Shen, and Wolfe, 1998; Zhu, Fung, and He, 2008). For any interval $[a, b]$, let $a = \kappa_{1-r} = \cdots = \kappa_0 < \kappa_1 < \cdots < \kappa_{J+1} = \cdots = \kappa_{J+r} = b$ be

a sequence of knots, normalized B-spline functions (Schumaker, 1981; Zhou, Shen, and Wolfe, 1998) of order r are defined as

$$B_j(t) = (\kappa_j - \kappa_{j-r})[\kappa_{j-r}, \ldots, \kappa_j](\kappa - t)_+^{r-1}, \quad j = 1, \ldots, K,$$

where $K = J + r$, $[\kappa_{j-r}, \ldots, \kappa_j]g(\kappa)$ is the rth order divided difference of the function $g(\kappa)$ on $\kappa_{j-r}, \ldots, \kappa_j$ and $(x)_+ = \max(x, 0)$. We collect B-spline basis functions in a K-dimensional vector as $\mathbf{B}(t) = \{B_1(t), \ldots, B_K(t)\}^{\mathrm{T}}$. The B-spline functions can also be evaluated using the recursive formula provided in Appendix D.

Suppose the study is conducted within a compact time interval $\mathcal{T}$ so that any subject, who has not failed or dropped out by the end of $\mathcal{T}$, is automatically right censored. Define the class of Hölder continuous functions on $\mathcal{T}$ as

$$C^{r,a}(\mathcal{T}) = \{f : \sup_{t_1,t_2\in\mathcal{T}} |f^{(r)}(t_1) - f^{(r)}(t_2)|/|t_1 - t_2|^a < \infty\} \tag{2.8}$$

for some nonnegative integer r and some $a > 0$, where $f^{(r)}$ is the rth derivative of f and a is called the Hölder exponent. Hölder continuity is a strong form of continuity for functions. If $f \in C^{r,a}(\mathcal{T})$, not only is $f^{(r)}$ continuous over $\mathcal{T}$, but $f^{(r)}(t_1) \to f^{(r)}(t_2)$ no slower than the rate of $|t_1 - t_2|^a$ as $t_1 \to t_2$. Define the L_∞ norm of a function f on $\mathcal{T}$ as $\|f\|_\infty = \sup_{t\in\mathcal{T}} |f(t)|$.

Recall that $\nu(t) = \log\{\lambda(t)\}$ and $\nu_{\mathrm{C}}(t) = \log\{\lambda_{\mathrm{C}}(t)\}$ are the log baseline hazard functions for the failure time and censoring time respectively, we model $\nu(t)$ and $\nu_{\mathrm{C}}(t)$ as splines to guarantee that estimators of $\lambda(t)$ and $\lambda_{\mathrm{C}}(t)$ are non-negative. We assume that the true log hazard functions are $\nu_0 \in C_1^{r_1,a_1}$ and $\nu_{\mathrm{C},0} \in C^{r_2,a_2}(\mathcal{T})$ for some r_1, r_2, a_1, $a_2 > 0$. Let $\mathbf{B}_1(t)$ and $\mathbf{B}_2(t)$ be K_1- and K_2-dimensional vectors of normalized B-spline basis functions of orders r_1 and r_2 with interior knots equally placed in $\mathcal{T}$. We allow K_1 and K_2 to increase slowly with the sample size, and define $\widetilde{\mathbf{B}}_1(t) = K_1^{1/2}\mathbf{B}_1(t)$ and $\widetilde{\mathbf{B}}_2(t) = K_2^{1/2}\mathbf{B}_2(t)$ for the convenience of asymptotic analysis. Under the Hölder continuity assumption above, there exist the best spline approximations to the true log hazard functions denoted by $\nu^*(t) = \widetilde{\mathbf{B}}_1^{\mathrm{T}}(t)\boldsymbol{\gamma}_1^*$ and $\nu_{\mathrm{C}}^*(t) = \widetilde{\mathbf{B}}_2^{\mathrm{T}}(t)\boldsymbol{\gamma}_2^*$ for some spline

coefficient vectors $\boldsymbol{\gamma}_1^*$ and $\boldsymbol{\gamma}_2^*$ such that $\|\nu^* - \nu_0\|_\infty = O(K_1^{-r_1})$ and $\|\nu_{\mathrm{C}}^* - \nu_{\mathrm{C},0}\|_\infty = O(K_2^{-r_2})$ (Schumaker, 1981).

Now we turn to the restricted model with ν and ν_{C} parameterized as $\nu(t) = \widetilde{\mathbf{B}}_1^{\mathrm{T}}(t)\boldsymbol{\gamma}_1$ and $\nu_{\mathrm{C}}(t) = \widetilde{\mathbf{B}}_2^{\mathrm{T}}(t)\boldsymbol{\gamma}_2$. The "true" parameters under this model are collected in $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_{10}^{\mathrm{T}}, \boldsymbol{\theta}_2^{*\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\theta}_{10}$ is the true value of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2^* = (\boldsymbol{\gamma}_1^{*\mathrm{T}}, \boldsymbol{\gamma}_2^{*\mathrm{T}})^{\mathrm{T}}$ contains the coefficients of best possible spline approximations. When ν and ν_{C} are nonparametric and belong to the Hölder class defined above, the estimation equations in (2.7) are slightly biased with

$$\mathrm{E}\{\mathbf{S}_n(\boldsymbol{\theta}^*)\} = \widetilde{O}(\max_{i=1,2} K_i^{-r_i}), \tag{2.9}$$

where $\widetilde{O}(\cdot)$ denotes the order uniformly for all entries of a vector or a matrix.

To develop the asymptotic theory for the proposed estimator bases on regression spline approximation, we first introduce some notations and conditions. We use $\|\cdot\|$ to denote the L^2 norm of either a vector or a function: $\|\boldsymbol{a}\| = (\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a})^{1/2}$ for a vector $\boldsymbol{a}$ and $\|f\| = \{\int_{\mathcal{T}} f^2(t)dt\}^{1/2}$ for a function f in the Hölder class. Denote the parameter in (2.2) with nonparametric log-hazard functions by $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \nu, \nu_{\mathrm{C}})$, which is an element in a Banach space with the norm $\|\boldsymbol{\Theta}\| = (\|\boldsymbol{\theta}_1\|^2 + \|\nu\|^2 + \|\nu_{\mathrm{C}}\|^2)^{1/2}$. Let $\ell(\boldsymbol{\Theta}) = \log\{p(\mathbf{D} \mid \boldsymbol{\Theta})\}$ be log-likelihood of data $\mathbf{D}$, define its Gâteaux derivative along the direction $\boldsymbol{\Theta}^\dagger = (\boldsymbol{\theta}_1^\dagger, \nu^\dagger, \nu_{\mathrm{C}}^\dagger)$ as

$$d\ell(\boldsymbol{\Theta}; \boldsymbol{\Theta}^\dagger) = \lim_{t \to 0} \{\ell(\boldsymbol{\Theta} + t\boldsymbol{\Theta}^\dagger) - \ell(\boldsymbol{\Theta})\}/t.$$

Denote the projection into the subspace orthogonal to $\Lambda^{\mathrm{R}*}$ by $\Pi_\perp^*(\cdot)$ and define

$$\mathbb{S}^*(\boldsymbol{\Theta}; \boldsymbol{\Theta}^\dagger) = \Pi_\perp^*[\mathrm{E}^*\{d\ell(\boldsymbol{\Theta}; \boldsymbol{\Theta}^\dagger) \mid \mathbf{O}\}]. \tag{2.10}$$

Let $\mathbb{S}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^\dagger)$ be a special case of $\mathbb{S}^*(\boldsymbol{\Theta}; \boldsymbol{\Theta}^\dagger)$ when $p(\boldsymbol{x} \mid \mathbf{z})$ is correctly specified.

We will make the following assumptions:

Assumptions.

(1) Denote the true parameter by $\boldsymbol{\Theta}_0 = (\boldsymbol{\theta}_{10}, \nu_0, \nu_{\mathrm{C},0})$ and assume that there exist some positive constants C_0, C_1 and C_2 such that

$$C_1 \leq E\{\mathbb{S}^*(\boldsymbol{\Theta}; \boldsymbol{\Theta}^\dagger)\mathbb{S}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^\dagger)\}/\|\boldsymbol{\Theta}^\dagger\|^2 \leq C_2,$$

for all $\boldsymbol{\Theta}^{\dagger}$ and $\boldsymbol{\Theta}$ with $\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_0\| \leq C_0$.

(2) Assume $K_i = O(n^{l_i})$ for $i = 1, 2$, with $\max_{i=1,2}(l_i) < 1/3$ and $\min_{i=1,2}(l_i r_i) > 1/2$.

Remark 4 Assumption (1) imposes a nondegeneracy and boundedness condition for the semiparametric efficient score (Tsiatis and Ma, 2004; Shen, 1997). Assumption (2) specifies the rate of growth of the dimension of the spline spaces relative to the sample size. It means that if higher order B-spline basis functions are used, fewer functions are required. It also implies that $r_i \geq 2, i = 1, 2$, so linear or higher order spline functions should be used.

Under these assumptions, we present the theoretical properties of the proposed estimators based on spline approximations. Proofs of the following theorems are provided in the Appendix.

Theorem 1 *The estimating equations in (2.7) yield a sequence of consistent solutions with* $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| = O_p(\delta_n)$, *where* $\delta_n = (K/n)^{1/2} + \max_{i=1,2} K_i^{-r_i+1/2}$.

Theorem 2 *Suppose* $\boldsymbol{\Gamma}$ *and* $\boldsymbol{\Sigma}$ *defined in (2.17) exist,*

$$n^{1/2}(\widehat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10}) \to \mathit{Normal}(\mathbf{0}, \boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1})$$

in distribution. If the conjectured model $p^*(\boldsymbol{x} \mid \mathbf{z})$ *is specified correctly, the asymptotic covariance matrix can be further simplified as* $\boldsymbol{\Sigma}^{-1}$.

Theorem 3 *The proposed semiparametric estimator* $\widehat{\boldsymbol{\theta}}_{1n}$ *is locally efficient. That is, the estimator is semiparametrically efficient if* $p^*(\boldsymbol{x} \mid \mathbf{z})$ *is specified correctly, and is still consistent if* $p^*(\boldsymbol{x} \mid \mathbf{z})$ *is specified incorrectly.*

Theorem 1 suggests that there exist a sequence of solutions to the estimating equation that converge to the true parameter. However, as in all estimating equation literature, the uniqueness of the solution is hard to show and often not guaranteed. The result

in Theorem 1 does not guarantee uniqueness of the solution to our estimating equation either in finite sample or in large sample limit. Carroll et al. (2006) recommend to solve the estimating equation with multiple initial values and, if distinct solutions are found, choose the one that is the closest to the naive estimator. In our simulation studies, we have never encountered multiple root problems with our estimator.

Recall that the parameter vector $\boldsymbol{\theta}$ contains both the parametric component θ_1 and the spline coefficients $\boldsymbol{\theta}_2$. For both theoretical derivation (see our proof of Theorem 2) and practical calculation, we need to calculate the asymptotic covariance of $\widehat{\boldsymbol{\theta}}_n$. As the dimension of spline basis functions increases, this covariance matrix also has an increasing dimension. In Theorem 2, we focus on the asymptotic distribution of the parametric part, the dimension of which is finite, and thus can establish a root-n convergence rate. The asymptotic covariance of $\widehat{\theta}_{1n}$ is the upper left sub-matrix of a much bigger covariance matrix with an increasing dimension. In calculating this sub-matrix, we define $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ as limits of various matrices. They have complicated forms that are hard to use directly. In practice, we estimate the covariance $\widehat{\boldsymbol{\theta}}_n$ directly using a sandwich formula (see Section 2.2.4 for more details), and the covariance of $\widehat{\boldsymbol{\theta}}_{1n}$ is extracted as a sub-matrix.

2.2.4 Implementation details

For given K_i and r_i, $i = 1, 2$, we use the following algorithm to solve $\widehat{\boldsymbol{\theta}}_n$ from the estimating equation (2.7) and obtain the estimate of its covariance matrix.

Algorithm.

(1) Choose an appropriate initial value for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$. Specifically, an initial value for $\boldsymbol{\theta}_1$ is obtained using the conditional score method (Tsiatis and Davidian, 2001; Song, Davidian, and Tsiatis, 2002). The initial value for $\boldsymbol{\theta}_2 = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$ is obtained through a naive maximum likelihood method (Kooperberg, Stone, and Truong, 1995) where we model $\nu(t)$ and $\nu_{\mathrm{C}}(t)$ as splines and replace X_i with W_i (or $\overline{W}_i$ if replicates of W_i are available).

(2) Solve $\widehat{\boldsymbol{\theta}}_n$ from (2.7) using a derivative-free optimization algorithm. For example, we use the function *lsqnonlin* in *MATLAB*. In order to solve (2.7), we first need to solve $\boldsymbol{h}(\mathbf{X}, \mathbf{Z})$ from the integral equation (2.5), where we use a simple approximation technique proposed by Tsiatis and Ma (2004). We discretize the value of $\mathbf{X}$ on m points, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, spread across the support of $\mathbf{X}$, so that the function $\boldsymbol{h}(\mathbf{X}, \mathbf{Z})$ becomes a q by m matrix $\boldsymbol{h}(\mathbf{Z}) = \{\boldsymbol{h}(\boldsymbol{x}_1, \mathbf{Z}), \ldots, \boldsymbol{h}(\boldsymbol{x}_m, \mathbf{Z})\}$ for any $\mathbf{Z}$. Then, the integral equation (2.5) is reduced to linear equations $\mathbf{A}(\mathbf{Z})\boldsymbol{h}^{\mathrm{T}}(\mathbf{Z}) = \boldsymbol{b}^{\mathrm{T}}(\mathbf{Z})$, where $\mathbf{A}(\mathbf{Z})$ and $\boldsymbol{b}(\mathbf{Z})$ are $m \times m$ and $q \times m$ matrices defined similarly as in equation (20) in Tsiatis and Ma (2004) but adapted to fit our setting.

(3) We estimate the covariance matrix of $\widehat{\boldsymbol{\theta}}_n$ by a sandwich formula similar to equation (18) in Tsiatis and Ma (2004)

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\theta}}_n) = \{\mathbf{J}_n(\widehat{\boldsymbol{\theta}}_n)\}^{-1}\mathbf{G}_n(\widehat{\boldsymbol{\theta}}_n)\{\mathbf{J}_n^{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_n)\}^{-1},$$

where $\mathbf{J}_n(\boldsymbol{\theta}) = -\partial/\partial\boldsymbol{\theta}\{\mathbf{S}_n(\boldsymbol{\theta})\}$ and $\mathbf{G}_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\mathbf{S}^*_{\mathrm{eff}}(\mathbf{O}_i, \boldsymbol{\theta})\mathbf{S}^{*\mathrm{T}}_{\mathrm{eff}}(\mathbf{O}_i, \boldsymbol{\theta})$. The derivative $\mathbf{J}_n(\boldsymbol{\theta})$ does not have an explicit form and is computed using numerical differentiation. The covariance of $\widehat{\boldsymbol{\theta}}_{1n}$ is obtained as the corresponding sub-matrix of $\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\theta}}_n)$.

In reality, the choice of spline basis functions, including the orders, numbers of knots and knots placement, is part of the model tuning. The optimal orders r_1 and r_2 depend on the smoothness of $\nu(t)$ and $\nu_{\mathrm{C}}(t)$, but in practice the results are least sensitive to these choices and cubic splines (twice continuously differentiable, piecewise cubic polynomial) are often employed (Kooperberg, Stone, and Truong, 1995). Suppose the study is conducted within a compact time interval $\mathcal{T} = [t_{\min}, t_{\max}]$, we place the interior knots with equal distance in $\mathcal{T}$. This choice is equivalent to using a global bandwidth in kernel regression and works well in our simulation studies and data analysis. Placing the knots adaptively is possible (Stone et al., 1997) but computationally much more intensive. With the orders and knots placement pre-determined, the performance of the method

is mainly governed by the numbers of spline basis K_1 and K_2. We choose these tuning parameters by minimizing the following Akaike information criteria (AIC)

$$\mathrm{AIC}(K_1, K_2) = -2\ell^*_{\mathbf{O}}(\widehat{\boldsymbol{\theta}}_n) + 2(p + K_1 + K_2),$$

where $\ell^*_{\mathbf{O}}(\cdot)$ denotes the log-likelihood for the observed data by integrating out $p^*(\boldsymbol{x} \mid \mathbf{z})$ and $\widehat{\boldsymbol{\theta}}_n$ is the estimator in the restricted model. Under the assumption that the censoring time is independent of the failure time and the covariates, it can be shown that $\ell^*_O(\boldsymbol{\theta})$ can be separated into two parts, $\ell^*_O(\boldsymbol{\theta}) = \ell^*_{O,1}(\boldsymbol{\theta}_1, \boldsymbol{\gamma}_1) + \ell^*_{O,2}(\boldsymbol{\gamma}_2)$, so the AIC can also be separated into two parts. Thus, K_1 and K_2 can be chosen separately. AIC works well in our numerical studies although other information criteria such as BIC (Kooperberg, Stone, and Truong, 1995) can also be used.

2.3 Simulation studies

2.3.1 Simulation 1: a linear Cox model

In the first simulation study, we considered a Cox model where the log hazard is linearly related to a covariate X, i.e. $\lambda(t \mid X) = \lambda_0(t)\exp(\beta X)$. The covariate X was generated from Normal(μ_X, σ_X^2) with $\mu_X = 1$ and $\sigma_X = 1$. Instead of observing X, we observed two independent copies of the surrogate $W^{[j]} = X + U^{[j]}, j = 1, 2$, where $U^{[j]}$ was generated independently from Normal$(0, \sigma_U^2)$ with $\sigma_U = 1$. We considered the baseline hazard functions $\lambda_0(t) = 1$, corresponding to an exponential distribution, and $\lambda_0(t) = t$, corresponding to a Weibull distribution. The true value of β was taken to be one. We considered both independent and dependent censoring mechanisms. Let b be a general constant which was used to achieve censoring rates of 30% and 60% in different scenarios. Under independent censoring, C was generated independently from a uniform distribution on $[0, b]$. Under the dependent censoring mechanism, C and Z are dependent, which is a departure from our assumption. Specifically, if $X > \mu_X$, C

follows a uniform distribution on $[0, b]$; otherwise, C follows a uniform distribution on $[0, 2b]$. We set the sample size to be $n = 500$ and repeated the simulation 1,000 times.

We used cubic splines to approximate the logarithms of $\lambda_0(t)$ and $\lambda_0^C(t)$. The numbers of spline basis functions K_1 and K_2 were selected by AIC as described in Section 2.4. The proposed estimator involves solving the integral equation (2.5), which requires specification of the density $p^*(x)$. We considered two choices of $p^*(x)$: a normal density with mean μ_X and variance σ_X^2 which corresponded to the truth and a uniform density on $[\mu_X - 3\sigma_X, \mu_X + 3\sigma_X]$ to assess the effect of misspecification. We used the method-of-moment estimators to estimate μ_X, σ_U^2, and σ_X^2, that is to say, $\widehat{\mu}_X = \sum_{i=1}^n \overline{W}_i/n$, $\widehat{\sigma}_U^2 = 2S_{\mathbf{V}}^2$, and $\widehat{\sigma}_X^2 = S_{\mathbf{W}}^2 - \widehat{\sigma}_U^2/2$, where $\overline{W}_i = (W_i^{[1]} + W_i^{[2]})/2$, $\overline{V}_i = (W_i^{[1]} - W_i^{[2]})/2$, and $S_{\mathbf{W}}^2$ and $S_{\mathbf{V}}^2$ are sample variances for $\overline{\mathbf{W}} = (\overline{W}_1, \ldots, \overline{W}_n)^{\mathrm{T}}$ and $\overline{\mathbf{V}} = (\overline{V}_1, \ldots, \overline{V}_n)^{\mathrm{T}}$ respectively. To solve equation (2.5), we adopted a simple discrete approximation for $p^*(x)$ following Tsiatis and Ma (2004). Specifically, we took X to be discrete with masses on 12 equally spaced points in $[\widehat{\mu}_X - 3\widehat{\sigma}_X, \widehat{\mu}_X + 3\widehat{\sigma}_X]$ with probabilities proportional to the density $p^*(x)$. All integrals involved were calculated by Gaussian quadratures using 100 nodes.

For comparison, we also considered various existing methods. The naive estimator maximizes the partial likelihood using $\overline{W}$ as the covariate. Two approximation methods were considered: the regression calibration (RC) estimator (Chapter 4 of Carroll et al. (2006)), where X is replaced by $E(X \mid \overline{W})$ in the partial likelihood, and the simulation and extrapolation (SIMEX) estimator (Chapter 5 of Carroll et al. (2006)) with a quadratic extrapolation function. A joint Gaussian distribution was assumed for X and $\overline{W}$ in regression calibration. The nonparametric correction (NPC) estimator of Huang and Wang (2000) and the conditional score (CS) estimator of Tsiatis and Davidian (2001) were also included for comparison.

The results under independent censoring are summarized in Table 2.1, where we report the bias, the empirical standard deviation, the mean estimated standard error

and the coverage rate of a 95% confidence interval for various estimators of β. As we can see, the naive estimator is severely biased; both approximation methods, i.e. RC and SIMEX, are still considerably biased; the NPC and CS estimators are consistent estimators and yield smaller biases than the approximation methods. In all cases shown in the table, our proposed estimators yield the smallest biases among all and have much smaller standard deviations than other consistent estimators (i.e. the NPC and CS) which illustrates the efficiency of our methods. For example, our methods are 77% to 89% more efficient than the CS estimator. The NPC method occasionally yields strange solutions and these outliers inflate the empirical bias and standard deviation. To be generous, we also report a cleaned version of NPC (labeled as NPC* in the table), where we remove 0.5% to 2% of outliers in different cases. Such failure to converge has never occurred to our estimators and even after removing these outliers the NPC still has much larger standard deviations than ours. In addition, both the NPC and CS methods significantly underestimate the standard errors, and our further simulations reveal that this underestimation of standard error can be severe when either the measurement error is large or the censoring rate is high. In contrast, the estimated standard errors of our methods do not show obvious underestimation. One striking finding of particular interest is that misspecification of $p(x)$ in our methods does not cause obvious loss of efficiency.

The results under dependent censoring are reported in Table 2.2. These results tell a rather similar story as Table 2.1. Our proposed method still vastly outperforms the competing methods, which also suggests that the proposed estimator is robust to mild violation of the independent censoring assumption.

2.3.2 Simulation 2: a quadratic Cox regression model

To assess the performance of the proposed estimator when $g(\cdot)$ in (2.1) is nonlinear, we considered the model $\lambda(t \mid X) = \lambda_0(t)\exp(\beta_1 X + \beta_2 X^2)$, where $\beta_1 = \beta_2 = -1$. We generated X from Normal$(-1, 1)$ and $W = X + U$ where $U \sim$ Normal$(0, \sigma_U^2)$ with

Table 2.1 Estimation results of β under Simulation 1 and independent censoring. The results are based on 1,000 replications.

		Exponential				Weibull			
Censoring	Estimator	Bias	ESD	ESE	Cov	Bias	ESD	ESE	Cov
30%	Naive	-428	51	49	0	-434	53	50	0
	RC	-139	88	74	518	-148	89	75	475
	SIMEX	-186	84	60	213	-192	87	61	183
	NPC	67	271	168	913	74	277	210	908
	NPC*	48	200	166	924	56	216	209	919
	CS	28	176	151	924	29	183	159	926
	Semi mis	12	128	121	932	14	140	127	929
	Semi true	11	128	119	930	16	134	128	939
60%	Naive	-396	65	63	0	-405	67	64	0
	RC	-90	108	95	786	-103	108	97	757
	SIMEX	-154	103	77	468	-162	106	78	450
	NPC	72	248	172	897	105	439	186	900
	NPC*	59	211	171	906	66	224	184	914
	CS	41	198	167	922	43	207	179	928
	Semi mis	21	149	141	957	25	154	148	953
	Semi true	21	148	139	954	23	152	145	949

Naive, the naive estimator; RC, regression calibration; SIMEX, the simulation extrapolation estimator; NPC, the nonparametric correction estimator of Huang and Wang (2000); NPC*, the NPC estimator after removing some outiers; CS, the conditional score estimator; Semi mis and Semi true, the locally efficient semiparametric estimators under the misspecified and true distribution of X, respectively; Bias, the empirical bias ($\times 10^3$); ESD, the empirical standard deviation ($\times 10^3$); ESE, the mean estimated standard error ($\times 10^3$); Cov, the empirical coverage probability of a 95% Wald confidence interval ($\times 10^3$).

$\sigma_U = 1/5$. The rest of the simulation setting is similar to the setting described in Simulation 1.

The NPC estimator was proposed for the case where $g(\cdot)$ is linear and therefore was not included for comparison. Since no complete and sufficient statistic exists under this model, the CS approach cannot be applied directly. Instead, we considered the approximated conditional score method of Song, Davidian, and Tsiatis (2002), which is based on a linear approximation using the delta method. The results for $\lambda_0(t) = t$ under independent censoring are summarized in Table 2.3. The Naive estimator is serenely

Table 2.2 Estimation results of β under Simulation 1 and dependent censoring. The results are based on 1,000 replications.

		Exponential				Weibull			
Censoring	Estimator	Bias	ESD	ESE	Cov	Bias	ESD	ESE	Cov
30%	Naive	-432	52	50	0	-440	53	50	0
	RC	-143	88	75	492	-154	89	76	441
	SIMEX	-187	85	60	206	-195	87	61	190
	NPC	72	264	166	918	88	287	178	907
	NPC*	60	206	166	924	66	216	177	921
	CS	35	178	153	938	36	190	163	937
	Semi mis	6	130	119	934	8	137	128	927
	Semi true	7	131	119	935	28	149	141	944
60%	Naive	-403	68	64	0	-419	69	66	0
	RC	-101	111	96	760	-126	112	99	696
	SIMEX	-159	107	78	470	-173	110	81	423
	NPC	73	337	168	887	95	418	185	880
	NPC*	53	216	167	896	61	231	183	893
	CS	37	200	163	924	39	210	178	935
	Semi mis	4	147	132	929	3	152	139	938
	Semi true	3	149	134	924	-9	155	143	933

Note: The layout of the table is similar to Table 2.1.

biased. The proposed estimator is robust against misspecification of $p(x)$ and has smaller bias than RC and approximated CS. The SIMEX estimator works surprisingly well under this particular setting and is comparable to our method.

2.4 Analysis of AIDS clinical trial data

We applied the proposed method to data from AIDS Clinical Trials Group (ACTG) 175 (Hammer et al., 1996) to assess the effects of antiretroviral therapies and baseline CD4 count on the time to AIDS or death in antiretroviral-naive patients. Four therapies were investigated in the ACTG 175 clinical trial, and previous studies (Hammer et al., 1996; Huang and Wang, 2000) found that the therapy using zidovudine alone is inferior to the other three therapies while the effects of the other three therapies are rather similar. Following Song and Huang (2005), we considered two treatment groups in our analysis,

Table 2.3 Results of the Simulation 2 based on 1,000 replications

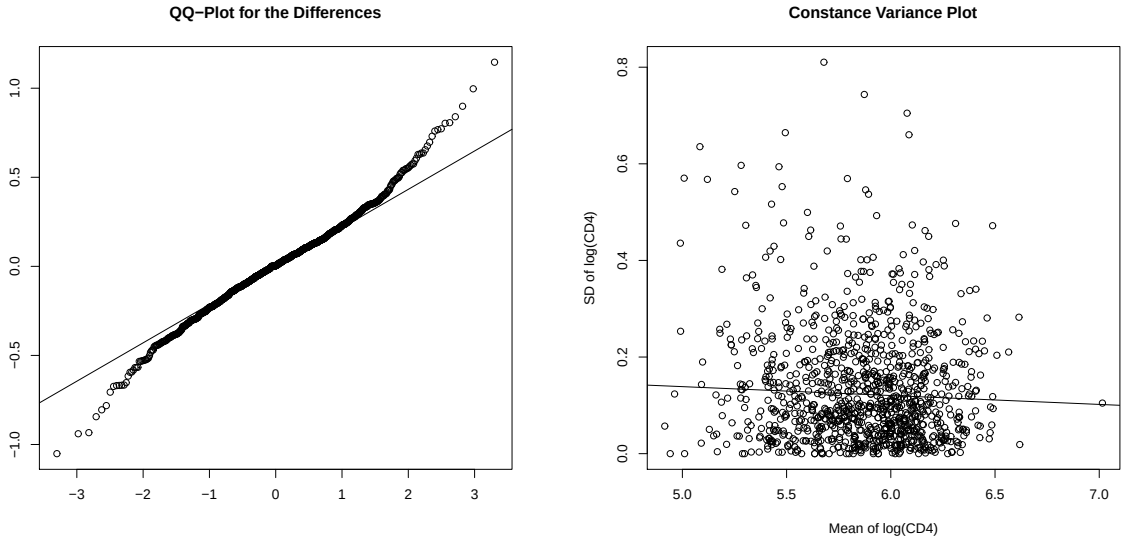
		β_1				β_2			
Censoring	Estimator	Bias	ESD	ESE	Cov	Bias	ESD	ESE	Cov
30%	Naive	181	124	124	662	142	79	76	507
	RC	76	136	134	890	71	86	82	832
	SIMEX	12	154	144	926	5	103	91	906
	CS*	-53	175	176	962	-46	120	123	974
	Semi mis	33	151	149	934	22	98	95	927
	Semi true	26	152	149	934	17	99	94	930
60%	Naive	144	189	190	846	117	119	116	781
	RC	35	206	206	941	44	129	126	908
	SIMEX	-15	225	218	950	-8	146	137	933
	CS*	-82	253	254	964	-59	168	168	968
	Semi mis	-9	225	221	955	-5	144	140	945
	Semi true	-15	227	223	952	-9	146	141	943

Note: The layout of the table is similar to Table 2.1. CS* is the approximated conditional score estimator of Song, Davidian, and Tsiatis (2002).

zidovudine alone and the combination of the other three therapies. It is well known that CD4 measurements were subject to substantial measurement errors. Among the 1067 antiretroviral-naive patients in the study, 1036 had two CD4 measurements within 3 weeks of randomization and 31 had only one CD4 measurement. The censoring rate for the time to AIDS or death was 91%.

Following the literature, we used the log transformed CD4 count as a covariate in the proportional hazards model. It is well-known that the CD4 count measures were subject to a significant amount of measurement error. We used the graphical tools described in Carroll et al. (2006) to check various assumptions on the measurement error. In the left panel of Figure 2.1, we show the normal Q-Q plot of the differences between replicates of log(CD4) within the same subject. The plot indicates that the measurement error exhibits slightly heavy tails on both sides, which is a mild deviation from the Gaussian assumption. In the right panel of Figure 2.1, we also plot the standard deviation of log(CD4) within a subject against the mean to check on the constant variance assumption. The regression line in this plot was fitted using the robust regression function

rlm in the *MASS* package in *R*. The estimated slope is -0.0184 with a p-value of 0.067, and therefore there is no clear violation of the assumption that the variance of measurement error is a constant. The estimated standard deviation is 0.182 for the measurement error and 0.276 for the true underlying log(CD4).



(a) (b)

Figure 2.1 Model checking for the AIDS clinical trial data. (a) is the normal QQ plot for the differences between replicates of log(CD4); (b) is the scatter plot of the standard deviation of log(CD4) within a subject versus the mean log(CD4) within the subject.

We considered a proportional hazards model with two covariates, the error-prone log(CD4) and an error-free treatment indicator which is zero for the zidovudine alone therapy and one for the other therapies. For the proposed locally efficient estimator, we considered two distributions for log(CD4), a normal distribution and a uniform distribution. In Table 2.4, we compare the estimation results based on various estimators considered in the simulation studies. In addition to the parameter estimates, we provide two versions of standard errors estimated by either a sandwich formula or a bootstrap procedure with 1,000 bootstrap samples. As suggested by our simulation results, the sandwich formula can sometimes underestimate the true standard deviation, especially

for the NPC and CS estimators. The bootstrap standard error, even though is computationally more expensive, is considered to be more reliable for a fair comparison among the different methods. Both the NPC and CS estimators are based on estimating equations, and the consistency of bootstrap for these estimators follows from the general theory by Chatterjee and Bose (2005) for bootstrapping estimating equations. Our estimator is based on a semiparametric estimating equation as we allow the number of splines and dimension of the estimating equation to diverge to infinity. We expect the consistency of bootstrap for our estimator to follow from similar arguments as in Chatterjee and Bose (2005), but a rigorous justification is out of the scope of this paper.

As we can see from Table 2.4, the treatment effect is not affected much by measurement error in CD4, since the naive estimate of the treatment effect is rather similar to the CS and the proposed estimators. The RC, SIMEX and NPC methods, however, seem to have overdone with bias-correction for the treatment effect. The naive estimator for the coefficient of log(CD4), on the other hand, is significantly attenuated compared with other estimators. All methods that take into account the measurement error provide similar estimates for log(CD4) except for the SIMEX which still shows a small degree of attenuation. The proposed estimator under either Gaussian or uniform assumption for log(CD4) has a smaller bootstrap standard error than other consistent estimators (i.e. the CS and NPC), indicating some efficiency gain using our method.

2.5 Discussion

We propose a class of locally efficient semiparametric estimators for the proportional hazards models with covariates contaminated with measurement error. Compared with competing methods, the proposed estimator is robust against misspecification of the distribution of the true covariate and is semiparametrically efficient if this underlying distribution is correctly specified. We allow the effect of the error-prone variable on the

Table 2.4 Results for the ACTG 175 data

	log(CD4)			Treatment		
Estimator	Est	SW SE	Boot SE	Est	SW SE	Boot SE
Naive	-1.879	0.327	0.366	-0.569	0.220	0.233
RC	-2.286	0.434	0.414	-0.675	0.216	0.226
SIMEX	-2.198	0.421	0.389	-0.681	0.219	0.227
NPC	-2.261	0.446	0.436	-0.670	0.222	0.232
CS	-2.262	0.392	0.445	-0.575	0.218	0.235
Semi uniform	-2.286	0.417	0.428	-0.585	0.225	0.237
Semi normal	-2.253	0.411	0.423	-0.580	0.224	0.236

Note: Naive, the naive estimator; RC, the regression calibration method; SIMEX, the simulation extrapolation estimator; NPC, the nonparametric correction estimator of Huang and Wang (2000); CS, the conditional score estimator; Semi uniform and Semi normal, the proposed semiparametric estimators when log(CD4) is assumed to be uniformly and normally distributed, respectively; Est, estimate; SW SE, standard error estimated using a sandwich formula; Boot SE, the bootstrap standard error based on 1,000 bootstrap replications.

failure time to have a very general parametric form, under which a sufficient statistic for the true covariate usually does not exist. The likelihood function of the proportion hazards model involves the integral of the baseline hazard function, which makes a local estimating equation approach like the one proposed by Ma and Carroll (2006) difficult to implement. We circumvent this difficulty using spline approximations. Our numerical studies show that our method vastly outperforms competing methods.

The efficiency loss for the proposed estimators under misspecified $p^*(\boldsymbol{x} \mid \mathbf{z})$ is unknown. However, the simulation results in this paper, as well as those in Tsiatis and Ma (2004) and Ma and Carroll (2006), show that the loss is virtually negligible even if the misspecification of $p^*(\boldsymbol{x} \mid \mathbf{z})$ is severe. In practice, it is still advisable to propose a proper specification of $p(\boldsymbol{x} \mid \mathbf{z})$ using some graphical tools. When the measurement error is very large, a nonparametric estimate of $p(\boldsymbol{x} \mid \mathbf{z})$ using deconvolution methods can also be employed (Delaigle, Hall, and Meister, 2008).

Our method is developed under the independent censoring assumption, but can be extended easily to situations where censoring time also depends on the covariates. In

those cases, another proportional hazards model can be used to characterize the relationship between the censoring time and the covariates, $\lambda_{i,\mathrm{C}}(t) = \lambda_{\mathrm{C}}(t)\exp\{g_{\mathrm{C}}(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}_{1,\mathrm{C}})\}$, where $g_{\mathrm{C}}(\cdot)$ is a known function of $\mathbf{X}_i$ and $\mathbf{Z}_i$ with a parameter $\boldsymbol{\theta}_{1,\mathrm{C}}$. We can redefine the parameter of interest to be a combination of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_{1,\mathrm{C}}$. Then a semiparametric estimator can be constructed similarly and the locally efficient property of the estimator still holds. Under this setting, we can estimate not only the relationship between the failure time and the covariates but also the censoring mechanism.

2.6 Acknowledgments

The research of Li and Xu was partially supported by the US National Science Foundation (DMS-1314118). The research of Song was partially supported by US National Science Foundation(DMS-1106816) and US National Institute of Health (R01ES017030).

2.7 Appendix: technical proofs

Appendix A: proof of theorem 1

We use the notation $\widetilde{O}_p(\cdot)$ and $\widetilde{o}_p(\cdot)$ to denote the element-wise $O_p(\cdot)$ and $o_p(\cdot)$ rates of a vector or a matrix. For any real valued matrix $\mathbf{A}$, define its spectral norm as $\|\mathbf{A}\| = \max_{\|\boldsymbol{x}\|\neq 0}\|\mathbf{A}\boldsymbol{x}\|/\|\boldsymbol{x}\|$ and its Frobenius norm as $\|\mathbf{A}\|_{\mathrm{F}} = \{\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A})\}^{1/2}$. Put $\mathbf{I}_n(\boldsymbol{\theta}) = E\{\mathbf{J}_n(\boldsymbol{\theta})\} = E\{-\partial/\partial\boldsymbol{\theta}\mathbf{S}^*_{\mathrm{eff}}(\mathbf{O}, \boldsymbol{\theta})\}$.

We use C, C_1 and C_2 as generic notation for positive constants. To show the existence of consistent solutions, we only need to verify the following condition (Ortega and Rheinboldt, 1970; Wang, 2011): for any $\epsilon > 0$, there exists a constant $\Delta > 0$ such that, for sufficiently large n,

$$\mathrm{pr}\{\sup_{\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^*\| = \Delta\delta_n} (\boldsymbol{\theta}_n - \boldsymbol{\theta}^*)^{\mathrm{T}}\mathbf{S}_n(\boldsymbol{\theta}_n) < 0\} \geq 1 - \epsilon. \tag{2.11}$$

For any $\boldsymbol{\Theta}_k = (\boldsymbol{\theta}_{1,k}, \nu_k, \nu_{\mathrm{C},k})$ with $\nu_k(t) = \widetilde{\mathbf{B}}_1^{\mathrm{T}}(t)\boldsymbol{\gamma}_k$ and $\nu_{\mathrm{C},k}(t) = \widetilde{\mathbf{B}}_2^{\mathrm{T}}\boldsymbol{\gamma}_{\mathrm{C},k}$, $k = 1, 2$, denote $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{1,k}^{\mathrm{T}}, \boldsymbol{\gamma}_k^{\mathrm{T}}, \boldsymbol{\gamma}_{\mathrm{C},k}^{\mathrm{T}})^{\mathrm{T}}$. By Lemma 6.1 of Zhou *et al.* (1998),

$$0 < C_1 \leq \|\nu_k\|^2/\|\boldsymbol{\gamma}_k\|^2, \|\nu_{\mathrm{C},k}\|^2/\|\boldsymbol{\gamma}_{\mathrm{C},k}\|^2 \leq C_2 < \infty,$$

and hence

$$0 < C_1 \leq \|\boldsymbol{\Theta}_k\|^2/\|\boldsymbol{\theta}_k\|^2 \leq C_2 < \infty, \quad k = 1, 2.$$

By the definition in (2.10), it is easy to see along the direction $\boldsymbol{\Theta}^\dagger$

$$\mathbb{S}^*(\boldsymbol{\Theta}_1; \boldsymbol{\Theta}_2) = \boldsymbol{\theta}_2^{\mathrm{T}}\mathbf{S}_{\mathrm{eff}}^*(\boldsymbol{\theta}_1).$$

For any $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq C\delta_n$, following similar lines of proof for equation (12) in Tsiatis and Ma (2004) while taking into account the asymptotic bias in $\mathbf{S}_{\mathrm{eff}}^*(\boldsymbol{\theta})$, we get

$$\mathbf{I}_n(\boldsymbol{\theta}) = E\{\mathbf{S}_{\mathrm{eff}}^*(\boldsymbol{\theta})\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}}(\boldsymbol{\theta})\} + \widetilde{O}(\delta_n). \tag{2.12}$$

By assumption (1), if $\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_0\| \leq C\delta_n$

$$0 < C_1 \leq \frac{\boldsymbol{\theta}_2^{\mathrm{T}}\mathbf{I}_n(\boldsymbol{\theta}_1)\boldsymbol{\theta}_2}{\|\boldsymbol{\theta}_2\|^2} \asymp \frac{E\{\mathbb{S}^*(\boldsymbol{\Theta}_1; \boldsymbol{\Theta}_2)\mathbb{S}(\boldsymbol{\Theta}_1; \boldsymbol{\Theta}_2)\}}{\|\boldsymbol{\Theta}_2\|^2} + O(K \times \delta_n) \leq C_2 < \infty, \tag{2.13}$$

where, for any sequences of positive constants $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means a_n/b_n is bounded away from both 0 and ∞. It is easy to see that $|\mathbf{J}_n(\boldsymbol{\theta}) - \mathbf{I}_n(\boldsymbol{\theta})| = \widetilde{O}(n^{-1/2})$ and hence $\|\mathbf{J}_n(\boldsymbol{\theta}) - \mathbf{I}_n(\boldsymbol{\theta})\|_{\mathrm{F}} = O(Kn^{-1/2})$. Therefore, as $n \to \infty$, with probability approaching 1

$$0 < C_1 \leq \frac{\boldsymbol{\theta}_2^{\mathrm{T}}\mathbf{J}_n(\boldsymbol{\theta}_1)\boldsymbol{\theta}_2}{\|\boldsymbol{\theta}_2\|^2} \leq C_2 < \infty \tag{2.14}$$

for any $\boldsymbol{\theta}_2$ provided that $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\| \leq C\delta_n$.

Finally, we verify (2.11). For any $\boldsymbol{\theta}_n$ satisfying $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^*\| = \Delta\delta_n$, $(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*)^{\mathrm{T}}\mathbf{S}_n(\boldsymbol{\theta}_n) := A_{n1} + A_{n2}$ with $A_{n1} = (\boldsymbol{\theta}_n - \boldsymbol{\theta}^*)^{\mathrm{T}}\mathbf{S}_n(\boldsymbol{\theta}^*)$ and $A_{n2} = -(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*)^{\mathrm{T}}\mathbf{J}_n(\bar{\boldsymbol{\theta}}_n)(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*)$ where $\bar{\boldsymbol{\theta}}_n$ is between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_n$. It is easy to see

$$E(|A_{n1}|) \leq \|\boldsymbol{\theta}_n - \boldsymbol{\theta}^*\|E\{\|\mathbf{S}_n(\boldsymbol{\theta}^*)\|\} \leq C\Delta\delta_n^2,$$

$$A_{n2} \leq -C_1\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^*\|^2 \leq -C_1\Delta^2\delta_n^2.$$

We can choose Δ large enough to make the probability of $\{(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*)^{\mathrm{T}}\mathbf{S}_n(\boldsymbol{\theta}_n) < 0\}$ approaching 1.

Appendix B: proof of theorem 2

By Taylor's expansion

$$0 = \mathbf{S}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{S}_n(\boldsymbol{\theta}^*) - \mathbf{J}_n(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) + \widetilde{O}_p(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|^2).$$

By Theorem 1 and assumption (2), it is easy to verify $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|^2 = o_p(n^{-1/2})$. Equation (2.14) also guarantees that $\mathbf{J}_n$ is non-singular and its eigenvalues are bounded away from 0. It follows that

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* = \mathbf{J}_n^{-1}(\boldsymbol{\theta}^*)\mathbf{S}_n(\boldsymbol{\theta}^*) + \widetilde{o}_p(n^{-1/2}). \tag{2.15}$$

Rewrite $\mathbf{S}_n$ and $\mathbf{S}^*_{\text{eff}}$ as $\mathbf{S}_n = (\mathbf{S}_{1n}^{\mathrm{T}}, \mathbf{S}_{2n}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{S}^*_{\text{eff}} = (\mathbf{S}_{1\text{eff}}^{*\mathrm{T}}, \mathbf{S}_{2\text{eff}}^{*\mathrm{T}})^{\mathrm{T}}$, according to the partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$. Write $\mathbf{J}_n$ and its inverse $\mathbf{J}_n^{-1}$ as

$$\mathbf{J}_n = \begin{pmatrix} \mathbf{J}_{11n} & \mathbf{J}_{12n} \\ \mathbf{J}_{21n} & \mathbf{J}_{22n} \end{pmatrix}, \qquad \mathbf{J}_n^{-1} = \begin{pmatrix} \mathbf{J}_n^{11} & \mathbf{J}_n^{12} \\ \mathbf{J}_n^{21} & \mathbf{J}_n^{22} \end{pmatrix}.$$

The first p equations in (2.15) becomes

$$\widehat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10} = \mathbf{J}_n^{11}(\boldsymbol{\theta}^*)\mathbf{S}_{1n}(\boldsymbol{\theta}^*) + \mathbf{J}_n^{12}(\boldsymbol{\theta}^*)\mathbf{S}_{2n}(\boldsymbol{\theta}^*) + \widetilde{o}_p(n^{-1/2}), \tag{2.16}$$

where $\mathbf{J}_n^{11} = (\mathbf{J}_{11n} - \mathbf{J}_{12n}\mathbf{J}_{22n}^{-1}\mathbf{J}_{21n})^{-1}$ and $\mathbf{J}_n^{12} = (\mathbf{J}_{11n} - \mathbf{J}_{12n}\mathbf{J}_{22n}^{-1}\mathbf{J}_{21n})^{-1}\mathbf{J}_{12n}\mathbf{J}_{22n}^{-1}$.

Define the partition of $\mathbf{I}_n(\boldsymbol{\theta})$ similarly as $\mathbf{J}_n(\boldsymbol{\theta})$, put $\boldsymbol{\Gamma}_n(\boldsymbol{\theta}^*) = (\mathbf{I}_{11n} - \mathbf{I}_{12n}\mathbf{I}_{22n}^{-1}\mathbf{I}_{21n})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$, $\boldsymbol{\Sigma}_n(\boldsymbol{\theta}^*) = \mathrm{cov}(\mathbf{S}^*_{1\text{eff}} + \mathbf{I}_{12n}\mathbf{I}_{22n}^{-1}\mathbf{S}^*_{2\text{eff}})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ and

$$\boldsymbol{\Gamma} = \lim_{n\to\infty} \boldsymbol{\Gamma}_n, \quad \boldsymbol{\Sigma} = \lim_{n\to\infty} \boldsymbol{\Sigma}_n. \tag{2.17}$$

It follows from (2.16) and (2.17) that

$$\begin{aligned} n^{1/2}(\widehat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10}) &= n^{1/2}\mathbf{J}_n^{11}(\boldsymbol{\theta}^*)\{\mathbf{S}_{1n}(\boldsymbol{\theta}^*) + \mathbf{J}_{12n}(\boldsymbol{\theta}^*)\mathbf{J}_{22n}^{-1}(\boldsymbol{\theta}^*)\mathbf{S}_{2n}(\boldsymbol{\theta}^*)\} + \widetilde{o}_p(1) \\ &= n^{1/2}\boldsymbol{\Gamma}^{-1}\{\mathbf{S}_{1n}(\boldsymbol{\theta}^*) + \mathbf{I}_{12n}(\boldsymbol{\theta}^*)\mathbf{I}_{22n}^{-1}(\boldsymbol{\theta}^*)\mathbf{S}_{2n}(\boldsymbol{\theta}^*)\} \\ &\quad + \mathbf{R}_1 + \mathbf{R}_2 + \widetilde{o}_p(1), \end{aligned} \tag{2.18}$$

where

$$\begin{aligned}\mathbf{R}_1 &= n^{1/2}\mathbf{J}_n^{11}(\boldsymbol{\theta}^*)\{\mathbf{J}_{12n}(\boldsymbol{\theta}^*)\mathbf{J}_{22n}^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_{12n}(\boldsymbol{\theta}^*)\mathbf{I}_{22n}^{-1}(\boldsymbol{\theta}^*)\}\mathbf{S}_{2n}(\boldsymbol{\theta}^*),\\ \mathbf{R}_2 &= n^{1/2}\{\mathbf{J}_n^{11}(\boldsymbol{\theta}^*) - \boldsymbol{\Gamma}^{-1}\}\{\mathbf{S}_{1n}(\boldsymbol{\theta}^*) + \mathbf{I}_{12n}(\boldsymbol{\theta}^*)\mathbf{I}_{22n}^{-1}(\boldsymbol{\theta}^*)\mathbf{S}_{2n}(\boldsymbol{\theta}^*)\}.\end{aligned}$$

We first calculate the rate of $\mathbf{R}_1$. It is easy to see $\|\mathbf{J}_n^{11}\|_F = O_p(1)$ and $\mathbf{J}_{12n}\mathbf{J}_{22n}^{-1} - \mathbf{I}_{12n}\mathbf{I}_{22n}^{-1} = (\mathbf{J}_{12n} - \mathbf{I}_{12n})\mathbf{J}_{22n}^{-1} + \mathbf{I}_{12n}(\mathbf{J}_{22n}^{-1} - \mathbf{I}_{22n}^{-1})$. By straightforward calculations, $\|(\mathbf{J}_{12n} - \mathbf{I}_{12n})\mathbf{J}_{22n}^{-1}\|_{\mathrm{F}} \le \|\mathbf{J}_{12n} - \mathbf{I}_{12n}\|_{\mathrm{F}}\|\mathbf{J}_{22n}^{-1}\| = O_p\{(K/n)^{1/2}\} \times O_p(1) = o_p(1)$. For the second term, $\|\mathbf{I}_{12n}\|_{\mathrm{F}} = O(K^{1/2})$ and $\|\mathbf{J}_{22n}^{-1} - \mathbf{I}_{22n}^{-1}\|_{\mathrm{F}} = \|\mathbf{I}_{22n}^{-1}(\mathbf{J}_{22n} - \mathbf{I}_{22n})\mathbf{I}_{22n}^{-1}\|_{\mathrm{F}} \times \{1 + o_p(1)\}$ (see Section 5.8 in Horn and Johnson (1985)). Equation (2.13) implies $\|\mathbf{I}_{22n}^{-1}\| = O(1)$, and hence $\|\mathbf{I}_{22n}^{-1}(\mathbf{J}_{22n} - \mathbf{I}_{22n})\mathbf{I}_{22n}^{-1}\|_{\mathrm{F}} = O_p(\|\mathbf{J}_{22n} - \mathbf{I}_{22n}\|_{\mathrm{F}})$. Entries of J_{22n} are partial derivatives with respect to spline coefficients. Because B-splines have compact supports, $\mathbf{J}_{22n}$ has a band matrix structure. In other word, there are only a fixed number of non-zero entries in each row of J_{22n}. Detailed calculations show $\|\mathbf{J}_{22n} - \mathbf{I}_{22n}\|_{\mathrm{F}}^2 = O_p(K/n)$. Therefore, $\|\mathbf{J}_{12n}\mathbf{J}_{22n}^{-1} - \mathbf{I}_{12n}\mathbf{I}_{22n}^{-1}\|_{\mathrm{F}} = O_p(Kn^{-1/2})$. By (2.9), $\|\mathbf{S}_{2n}(\boldsymbol{\theta}^*)\|^2 \le \|\mathbf{S}_{2n}(\boldsymbol{\theta}^*) - E\{\mathbf{S}_{2n}(\boldsymbol{\theta}^*)\}\|^2 + \|E\{\mathbf{S}_{n2}(\boldsymbol{\theta}^*)\}\|^2 = O_p(K/n + K\max_{i=1,2} K_i^{-2r_i})$. Under assumption (2), the results above lead to

$$\|\mathbf{R}_1\| = O_p(K^{3/2}n^{-1/2} + K^{3/2}\max_{i=1,2} K_i^{-r_i}) = o_p(1).$$

By similar calculations as above, we have $\|(\mathbf{J}_{11n} - \mathbf{J}_{12n}\mathbf{J}_{22n}^{-1}\mathbf{J}_{21n}) - \boldsymbol{\Gamma}_n\|_{\mathrm{F}} \le \|\mathbf{J}_{11n} - \mathbf{I}_{11n}\|_{\mathrm{F}} + \|(\mathbf{J}_{12n} - \mathbf{I}_{12n})\mathbf{J}_{22n}^{-1}\mathbf{J}_{21n}\|_{\mathrm{F}} + \|\mathbf{I}_{12n}(\mathbf{J}_{22n}^{-1}\mathbf{J}_{21n} - \mathbf{I}_{22n}^{-1}\mathbf{I}_{21n})\|_{\mathrm{F}} = O_p(n^{-1/2}) + O_p(Kn^{-1/2}) + O_p(K^{3/2}n^{-1/2}) = o_p(1)$. Since $\boldsymbol{\Gamma}_n^{-1} \to \boldsymbol{\Gamma}^{-1}$, $\|\mathbf{J}_n^{11}(\boldsymbol{\theta}^*) - \boldsymbol{\Gamma}^{-1}\|_{\mathrm{F}} = o_p(1)$ and

$$\|\mathbf{R}_2\| = o_p(n^{1/2}\|\mathbf{S}_{1n}(\boldsymbol{\theta}^*) + \mathbf{I}_{12n}(\boldsymbol{\theta}^*)\mathbf{I}_{22n}^{-1}(\boldsymbol{\theta}^*)\mathbf{S}_{2n}(\boldsymbol{\theta}^*)\|) = o_p(1).$$

Since $\mathbf{R}_1$ and $\mathbf{R}_2$ are higher order terms, (2.18) can be re-written as

$$n^{1/2}(\widehat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10}) = n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\Gamma}^{-1}\left\{\mathbf{S}_{1\mathrm{eff}}^*(O_i, \boldsymbol{\theta}^*) + \mathbf{I}_{12n}(\boldsymbol{\theta}^*)\mathbf{I}_{22n}^{-1}(\boldsymbol{\theta}^*)\mathbf{S}_{2\mathrm{eff}}^*(O_i, \boldsymbol{\theta}^*)\right\} + o_p(1). \tag{2.19}$$

It is easy to see $nvar(\widehat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10}) = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}_n\boldsymbol{\Gamma}^{-1} + o(1) \to \boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1}$, and the asymptotic normality in Theorem 2 follows from the central limit theorem.

When $p(\boldsymbol{x} \mid \mathbf{z})$ is correctly specified, similar to (2.12), we can show that

$$\mathbf{I}_n(\boldsymbol{\theta}^*) = E\left\{\mathbf{S}_{\text{eff}}(O, \boldsymbol{\theta}^*)\mathbf{S}_{\text{eff}}^{\mathrm{T}}(O, \boldsymbol{\theta}^*)\right\} + \widetilde{O}(\delta_n).$$

Then it is easy to verify $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}$.

Appendix C: proof of theorem 3

We only need to show $\widehat{\boldsymbol{\theta}}_{1n}$ is semiparametrically efficient when $p(\boldsymbol{x} \mid \mathbf{z})$ is correctly specified. By (2.19), $\widehat{\boldsymbol{\theta}}_{1n}$ is asymptotically linear with the influence function

$$\boldsymbol{\varphi}(\mathbf{O}) = \boldsymbol{\Gamma}^{-1}\left\{\mathbf{S}_{1\text{eff}}(\mathbf{O}, \boldsymbol{\theta}^*) + \mathbf{I}_{12n}(\boldsymbol{\theta}^*)\mathbf{I}_{22n}^{-1}(\boldsymbol{\theta}^*)\mathbf{S}_{2\text{eff}}(\mathbf{O}, \boldsymbol{\theta}^*)\right\}.$$

By our construction of the efficiency scores, $\mathbf{S}_{\text{eff}} \perp (\Lambda_1 + \Lambda_2)$ and hence $\boldsymbol{\varphi} \perp (\Lambda_1 + \Lambda_2)$. Therefore, to show semiparametric efficiency, it suffices to prove $\boldsymbol{\varphi} \perp (\Lambda_3 + \Lambda_4)$.

For any K-dimensional constant vector $\mathbf{v}$, (2.12) implies

$$\langle \boldsymbol{\varphi}, \mathbf{S}_{2\text{eff}}^{\mathrm{T}}\mathbf{v}\rangle = \boldsymbol{\Gamma}^{-1}\{E(\mathbf{S}_{1\text{eff}}\mathbf{S}_{2\text{eff}}^{\mathrm{T}}\mathbf{v}) - E(\mathbf{I}_{12n}\mathbf{I}_{22n}^{-1}\mathbf{S}_{2\text{eff}}\mathbf{S}_{2\text{eff}}^{\mathrm{T}}\mathbf{v})\} = O(\max_{i=1,2} K_i^{-r_i}). \tag{2.20}$$

Recall $\mathbf{S}_{2\text{eff}}(\mathbf{O}) = \mathbf{S}_{\boldsymbol{\theta}_2}(\mathbf{O}) - \Pi\{\mathbf{S}_{\boldsymbol{\theta}_2}(\mathbf{O}) \mid \Lambda_1 + \Lambda_2\}$, where $\mathbf{S}_{\boldsymbol{\theta}_2}(\mathbf{O}) = \{\mathbf{S}_{\boldsymbol{\gamma}_1}^{\mathrm{T}}(\mathbf{O}), \mathbf{S}_{\boldsymbol{\gamma}_2}^{\mathrm{T}}(\mathbf{O})\}^{\mathrm{T}}$,

$$\begin{aligned}
\mathbf{S}_{\boldsymbol{\gamma}_1}(\mathbf{O}) &= E\{\mathbf{S}_{\boldsymbol{\gamma}_1}(V, \Delta, \mathbf{X}, \mathbf{Z}) \mid \mathbf{O}\} = E\{\textstyle\int \widetilde{\mathbf{B}}_1(u)\, \mathrm{d}M(u, \mathbf{X}, \mathbf{Z}) \mid \mathbf{O}\}, \\
\mathbf{S}_{\boldsymbol{\gamma}_2}(\mathbf{O}) &= E\{\mathbf{S}_{\boldsymbol{\gamma}_2}(V, \Delta, \mathbf{X}, \mathbf{Z}) \mid \mathbf{O}\} = \textstyle\int \widetilde{\mathbf{B}}_2(u)\, \mathrm{d}M_{\mathrm{C}}(u).
\end{aligned}$$

An arbitrary element $h(\mathbf{O})$ in $\Lambda_3 + \Lambda_4$ can be expressed as

$$h(\mathbf{O}) = c_1 E\{\textstyle\int a_1(u)\, \mathrm{d}M(u, \mathbf{X}, \mathbf{Z}) \mid O\} + c_2 \int a_2(u)\, \mathrm{d}M_C(u)$$

for some integrable functions $a_1(\cdot)$, $a_2(\cdot)$ and constants c_1 and c_2. Because the Hölder class defined in (2.8) is dense in L^2 functional space, there exists a K_i-dimensional vector $\mathbf{v}_i$ such that $\|a_i(u) - \widetilde{\mathbf{B}}_i^{\mathrm{T}}(u)\mathbf{v}_i\| = O(K_i^{-r_i})$, $i = 1, 2$. Let $\mathbf{v} = (c_1\mathbf{v}_1^{\mathrm{T}}, c_2\mathbf{v}_2^{\mathrm{T}})^{\mathrm{T}}$, then

$$\begin{aligned}
h(\mathbf{O}) &= c_1 E\{\textstyle\int \widetilde{\mathbf{B}}_1^{\mathrm{T}}(u)\mathbf{v}_1\, \mathrm{d}M(u, \mathbf{X}, \mathbf{Z}) \mid \mathbf{O}\} + c_2 \int \widetilde{\mathbf{B}}_2^{\mathrm{T}}(u)\mathbf{v}_2\, \mathrm{d}M_C(u) + O(\max_{i=1,2} K_i^{-r_i}) \\
&= \mathbf{v}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{\theta}_2}(\mathbf{O}) + O(\max_{i=1,2} K_i^{-r_i}) \\
&= \mathbf{v}^{\mathrm{T}}\mathbf{S}_{2\text{eff}}(\mathbf{O}) + \mathbf{v}^{\mathrm{T}}\Pi\{\mathbf{S}_{\boldsymbol{\theta}_2}(\mathbf{O}) \mid \Lambda_1 + \Lambda_2\} + O(\max_{i=1,2} K_i^{-r_i}).
\end{aligned} \tag{2.21}$$

Equations (2.20) and (2.21) imply $\langle \boldsymbol{\varphi}(\mathbf{O}), h(\mathbf{O}) \rangle = O(\max_{i=1,2} K_i^{-r_i}) = o\left(n^{-1/2}\right)$ for any $h(\mathbf{O}) \in \Lambda_3 + \Lambda_4$. Consequently, $\boldsymbol{\varphi}(\mathbf{O})$ is an efficient influence function which completes the proof.

Appendix D: Recursive formula for B-splines

Let $a < \kappa_1 < \kappa_2 < \cdots < \kappa_J < b$ be the sequence of inner knots as defined in Section 2.2.3 and let $\kappa_{1-r} = \ldots = \kappa_0 = a$ and $\kappa_{J+1} = \ldots = \kappa_{J+r} = b$ be the boundary knots. To distinguish B-splines with different orders, denote $\{B_{j,k}(t); j = 1, \ldots, J+k\}$ as the collection of kth order B-splines defined on the same set of inner knots for $k = 1, \ldots, r$. The first order B-splines or the constant B-splines are defined as

$$B_{j,1}(t) = 1 \quad \text{for } t \in [\kappa_{j-1}, \kappa_j), \quad 0 \text{ otherwise}, \quad j = 1, \ldots, J+1.$$

Then the kth order B-splines can be evaluated using the following recursive formula

$$B_{j,k}(t) = \frac{t - \kappa_{j-k}}{\kappa_{j-1} - \kappa_{j-k}} B_{j-1,k-1}(t) + \frac{\kappa_j - t}{\kappa_j - \kappa_{j-k+1}} B_{j,k-1}(t), \quad j = 1, \ldots, J+k,$$

with the convention that $B_{0,k-1}(t) \equiv 0$, $B_{J+k,k-1}(t) \equiv 0$ and $0/0 = 0$.

CHAPTER 3. SEMIPARAMETRIC ESTIMATION FOR MEASUREMENT ERROR MODELS WITH VALIDATION DATA

A paper submitted to *Canadian Journal of Statistics*

Yuhang Xu[1], Jae-kwang Kim[2] Yehua Li[3]

Abstract

We consider regression problems where error-prone surrogates of the true predictors are collected in the primary data set while accurate measurements of the predictors are available only in a small validation data set. We propose a new class of semiparametric estimators for the regression coefficients based on expected estimating equations, where the relationship between the surrogates and the true predictors is modeled nonparametrically using a kernel smoother trained with the validation data. The new methods are developed under two different scenarios where the response variable is either observed or not observed in the validation data set. The proposed estimators have a natural connection with the fractional imputation method. They are consistent, asymptotically unbiased and normal in both scenarios. Our simulation studies show that the proposed estimators are superior to competitors in terms of bias and mean square error and are

[1]Primary researcher and author, Graduate student, Department of Statistics, Iowa State University.
[2]Author for correspondence, Professor, Department of Statistics, Iowa State University.
[3]Associate Professor, Department of Statistics, Iowa State University.

quite robust against the misspecification of the regression model and bandwidth selection. A real data application in the Korean Longitudinal Study of Aging is presented for illustration.

3.1 Introduction

In many scientific studies, surrogates of the predictors are collected because precise measurements of the true predictors are either unavailable or too expensive. Examples of such surrogate measurements include food questionnaires in nutrition studies (Kipnis et al., 2009), calibrated radiation dose in radiation epidemiology (Li, et al., 2007), and CD4 count measurements in HIV clinical trials (Xu, Li, and Song, 2016). In our motivating example detailed in Section 3.5, one of the key predictors, the body mass index (BMI), is calculated based on self-reported weight and height and is subject to measurement errors. It is well known that ignoring measurement errors in the surrogates may cause estimation bias (Carroll et al., 2006). In order to properly account for measurement errors, a small validation data set is collected in many studies including our motivating application in Section 3.5, where both the true predictor and the surrogate are measured. Such a validate data set can be used to calibrate the measurement error model.

We consider a general regression problem, where Y is the response variable and X is a vector of p_1 explanatory variables. The conditional density of Y given X is $f(y \mid x; \beta)$, where β is a q dimensional parameter vector of interest. In the primary data set, the true value of X is unavailable and a cheaper surrogate W is observed instead. The primary data set consists of N independent copies of (Y, W), denoted as $\{(Y_i, W_i)_{i=1}^{N}\}$. We consider two scenarios for the validation data set. In Scenario I, Y is not observed in the validation set, and the validation data set consists of $\{(X_i, W_i)_{i=N+1}^{N+n}\}$; in Scenario II, Y is also observed in the validation set and the validation data set consists of $\{(Y_i, X_i, W_i)_{i=N+1}^{N+n}\}$. We refer to these two types of validation data as Type I and

Type Ⅱ validation data in this paper. The sample size of the validation data set, n, is usually much smaller than that of the primary data set, N.

Many methods in the measurement error literature rely on strong parametric assumptions on the conditional distribution $f(W \mid X)$, which may not hold in practice. The most commonly used model is the classical measurement error model (Fuller, 1987; Carroll et al., 2006), where $W = X + \epsilon$ and $\epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2)$. When validation data are available, semiparametric methods with much relaxed assumptions on $f(W \mid X)$ have been proposed in the literature. Lee and Sepanski (1995) and Sepanski and Lee (1995) propose to estimate β by minimizing a nonlinear least squares $N^{-1}\sum_{i=1}^{N}[Y_i - E\{G(X_i, \beta) \mid W_i\}]^2$, where $G(X, \beta) = E(Y \mid X)$ and the conditional expectation $E(\cdot \mid W)$ is calibrated nonparametrically using the validation data. Wang and Rao (2002) and Stute, Xue, and Zhu (2007) propose to estimate β using empirical likelihood methods based on the same philosophy as the nonlinear least squares. These methods assume $E\{Y - G(X, \beta) \mid W\} = 0$ with probability 1 and ignore the fact that $\mathrm{Var}(Y \mid W; \beta)$ is usually heteroscedastic in nonlinear regression models, so they may suffer from loss of efficiency.

There has also been some literature on semiparametric methods based on the likelihood or score equation on $[Y \mid X]$ and a nonparametric measurement error model calibrated using the validation data. These methods include those by Carroll and Wand (1991), Pepe and Fleming (1991), Reilly and Pepe (1995), Wang and Wang (1997), Chatterjee and Chen (2007), and Wang and Yu (2007). However, these methods have their own limitations. Carroll and Wand (1991) and Wang and Wang (1997) limit their focus to logistic regression models; Pepe and Fleming (1991) consider a more general regression setting but require W to be categorical. Reilly and Pepe (1995) propose a mean-score method for a related missing data problem, but their method is only applicable when X and W are discrete variables. Chatterjee and Chen (2007) and Wang and Yu (2007) consider a general regression problem and allow W to be continuous, but their score-based estimators are not robust. In addition, the method proposed by Chatterjee and Chen

(2007) can only be applied to Type Ⅱ validation data and the method proposed in Wang and Yu (2007) is inconsistent in general.

Our main contribution lies in that we propose a new class of semiparametric estimators based on expected estimating equations (Wang and Pepe, 2000) that are consistent and efficient. The expected estimating equations we proposed are nonparametric extensions of the ones proposed by Wang and Pepe (2000). We calibrate the conditional expectation in the expected estimating equations nonparametrically using the validation data and thus do not require any parametric assumption on the measurement errors. Our methods are applicable to any nonlinear regression model, under either type of validation data mentioned above and we allow W to be a continuous variable. Furthermore, our methods allow the estimating equation to be more general than a correctly specified score function and hence have the benefit of robustness.

The rest of the paper is organized as follows. We describe our main methodology in Section 3.2, where we propose estimation procedures under both Type I and Type Ⅱ validation data and study their asymptotic properties. In Section 3.2.3, we also describe an extension of the methods to the case where there is an error-free covariate Z related to Y. In Section 3.3, we address some of the computation issues. Simulation studies are presented in Section 3.4 and a real data application in the Korean Longitudinal Study of Aging is provided in Section 3.5. We conclude the paper by discussions in Section 3.6. All technical details are collected in the Appendix, Section 3.8.

3.2 Methodology

3.2.1 Derivation of the methodology

One basic assumption of our methods is that the subjects in the validation set are drawn from the same population as the primary set such that the measurement error model is "transportable" between the two sets of data (Chap.2 of Carroll et al. (2006)). In

what follows, we use $f(\cdot)$ to denote a density function and $f(\cdot \mid \cdot)$ to denote a conditional density function. Let $U(\beta; x, y)$ be an unbiased estimating function for β such that $E\{U(\beta; X, Y)\} = 0$. An optimal choice for $U(\beta; x, y)$ is the score function $S(\beta; x, y) = \partial/\partial\beta\{\log f(y \mid x; \beta)\}$, but other choices can be made for the reason of robustness.

We first consider Scenario I described in the introduction where the validation data set consists of $\{(X_i, W_i)_{i=N+1}^{N+n}\}$. The only source of information for β is from the primary data, where X is missing. A consistent estimator of β can be obtained by solving

$$\sum_{i=1}^{N} E\{U(\beta; X_i, Y_i) \mid Y_i, W_i\} = 0 \tag{3.1}$$

with respect to β. The equations (3.1) are often called expected estimating equations (Wang and Pepe, 2000); when $U(\beta; x, y)$ is the score function, they are often called mean score equations. Under a commonly used surrogacy assumption

$$f(y \mid x, w; \beta) = f(y \mid x; \beta), \tag{3.2}$$

the conditional expectation in (3.1) can be written as

$$E\{U(\beta; X_i, Y_i) \mid Y_i, W_i\} = \frac{\int U(\beta; x_i, Y_i) f(Y_i \mid x_i; \beta) f(x_i \mid W_i)\, \mathrm{d}x_i}{\int f(Y_i \mid x_i; \beta) f(x_i \mid W_i)\, \mathrm{d}x_i}. \tag{3.3}$$

If we define $g(x, y; \beta) = U(\beta; x, y) f(y \mid x; \beta)$ and $\mu_g(y, w; \beta) = E\{g(X, y; \beta) \mid W = w\}$, then the integral in the numerator of (3.3) is $\mu_g(Y_i, W_i; \beta)$. Similarly, the integral in the denominator of (3.3) also can be written as $\mu_g(Y_i, W_i; \beta)$ if we define $g(x, y; \beta) = f(y \mid x; \beta)$. For any fixed y, $\mu_g(y, w; \beta)$ is a nonparametric mean function of $g(X, y; \beta)$ given $W = w$, which can be estimated using kernel regression based on the validation data.

Let $K(\cdot)$ be a kernel function of order k (Fan and Gijbels, 1996), and denote $K_h(\cdot) = K(\cdot/h)/h^{p_1}$, where h is a bandwidth. Note that $h \equiv h_n$ is a tuning parameter that depends on the sample size, and we suppress the subscript n for ease of exposition. A kernel regression estimator of $\mu_g(y, w; \beta)$ based on the validation data is

$$\widehat{\mu}_g(y, w; \beta) = \frac{\sum_{j=N+1}^{N+n} g(X_j, y; \beta) K_h(w - W_j)}{\sum_{j=N+1}^{N+n} K_h(w - W_j)}.$$

We now approximate the expected estimating function (3.3) by

$$\bar{U}_i(\beta) = \frac{\sum_{j=N+1}^{N+n} U(\beta; X_j, Y_i) f(Y_i \mid X_j; \beta) K_h(W_i - W_j)}{\sum_{j=N+1}^{N+n} f(Y_i \mid X_j; \beta) K_h(W_i - W_j)}, \tag{3.4}$$

and obtain the following approximate expected estimating equations

$$\sum_{i=1}^{N} \bar{U}_i(\beta) = 0. \tag{3.5}$$

Denote the solution of (3.5) as $\widehat{\beta}_{\text{I}}$, where the subscription emphasizes the type of validation data being used. More details on selecting the bandwidth h and solving equation (3.5) are provided in Section 3.3.1.

Under Scenario II, the validation data consist of $\{(Y_i, X_i, W_i)_{i=N+1}^{N+n}\}$, which provide information for β in addition to the primary data. We propose to combine this additional estimating equation with equation (3.5), which is based on the primary data set. The proposed estimator under Scenario II is $\widehat{\beta}_{\text{II}}$, which is the solution of

$$\sum_{i=1}^{N} \bar{U}_i(\beta) + \sum_{i=N+1}^{N+n} U(\beta; X_i, Y_i) = 0. \tag{3.6}$$

3.2.2 Asymptotic theory

We now study the large sample properties of the proposed estimators. We denote the ratio of the two sample sizes as $\lambda_N = N/n$ and assume that $\lambda_N \to \lambda < \infty$. In other words, both the primary sample and the validation sample increase at the same rate. Some additional regularity conditions are presented in Appendix. To facilitate further theoretical derivation, define

$$\begin{aligned}
d(\beta; y, w) &= \int U(\beta; x, y) f(y \mid x; \beta) f(x, w)\, \mathrm{d}x, \\
c(\beta; y, w) &= \int f(y \mid x; \beta) f(x, w)\, \mathrm{d}x,\ e(\beta; y, w) = d(\beta; y, w)/c(\beta; y, w), \\
\bar{d}(\beta; y, w) &= n^{-1} \textstyle\sum_{j=N+1}^{N+n} U(\beta; X_j, y) f(y \mid X_j; \beta) K_h(w - W_j), \\
\bar{c}(\beta; y, w) &= n^{-1} \textstyle\sum_{j=N+1}^{N+n} f(y \mid X_j; \beta) K_h(w - W_j),\ \bar{e}(\beta; y, w) = \bar{d}(\beta; y, w)/\bar{c}(\beta; y, w).
\end{aligned}$$

Using the notation above, the estimating function and its approximation (3.4) can be expressed as

$$E\left\{U(\beta; X_i, Y_i) \mid Y_i, W_i\right\} = e(\beta; Y_i, W_i), \quad \bar{U}_i(\beta) = \bar{e}(\beta; Y_i, W_i). \tag{3.7}$$

We first study the asymptotic properties of the estimator under Scenario I. Define $M(\beta; x, w) = M_1(\beta; x) - M_2(\beta; x, w)$, where $M_1(\beta; x) = E\{U(\beta; X, Y)|X = x\}$ and $M_2(\beta; x, w) = E\{e(\beta; Y, W)|X = x, W = w\}$. The following theorem shows that $\widehat{\beta}_{\text{I}}$, the solution of equation (3.5), is asymptotically unbiased and follows an asymptotic normal distribution. The proof of the theorem is provided in Appendix.

Theorem 1 *Under the conditions listed in Appendix, we have*

$$N^{1/2}(\widehat{\beta}_{\text{I}} - \beta) \xrightarrow{D} N(0, V_1),$$

where $V_1 = \Gamma_1^{-1}\Sigma_1\Gamma_1^{-1}$, $\Gamma_1 = A$, $\Sigma_1 = \Gamma_1 + \lambda B$, $A = var\{e(\beta; Y, W)\}$, *and* $B = var\{M(\beta; X, W)\}$.

Remark 1 The asymptotic covariance matrix V_1 can be written as $V_1 = A^{-1} + \lambda A^{-1}BA^{-1}$. By (3.7), the first part, A^{-1}, is the inverse Fisher information from the primary data set if $f(x \mid w)$ is known, and the second part is the price we pay for modeling $f(x \mid w)$ nonparametrically.

Remark 2 We can estimate V_1 by replacing A and B with their method-of-moment estimators

$$\widehat{A} = N^{-1}\textstyle\sum_{i=1}^{N} \bar{e}^{\otimes 2}(\widehat{\beta}_{\text{I}}; Y_i, W_i), \qquad \widehat{B} = n^{-1}\textstyle\sum_{i=N+1}^{N+n} \widehat{M}^{\otimes 2}(\widehat{\beta}_{\text{I}}; X_i, W_i), \tag{3.8}$$

where $v^{\otimes 2} = vv^{\text{T}}$ for a vector v, $\widehat{M}(\beta; x, w) = M_1(\beta; x) - \widehat{M}_2(\beta; x, w)$, and $\widehat{M}_2(\beta; x, w) = \int f(y \mid x; \beta)\bar{e}(\beta; y, w)\, \mathrm{d}y$ which is evaluated using numerical integration. One complication with this "plug-in" method is that the optimal bandwidth for estimating β may not be the optimal one for variance estimation (Fan and Gijbels, 1996). In our simulation

studies, we find that this plug-in variance estimator tends to underestimate V_1 slightly. For inference purposes, we suggest estimating V_1 directly using a nonparametric bootstrap method, which gives a closer approximation to the true standard error. Details of the bootstrap procedure are described in our data analysis in Section 3.5.

Remark 3 Having separate estimators for A and B, however, has some additional benefits. For example, one can use A and B estimated from a pilot study to determine the optimal sample ratio λ^* for a large-scale study. Suppose c_0 and c_1 are the costs to collect the information of a participant for the primary and validation data, respectively. Usually c_1 is much higher than c_0 because an accurate measurement on X is expensive. The total cost of the study is $c = Nc_0 + nc_1$, and the covariance of $\widehat{\beta}_{\text{I}}$ is approximately $V_1/N = A^{-1}/N + A^{-1}BA^{-1}/n$. The optimal sample ratio λ^* would provide the highest precision with a fixed budget or lowest cost with a fixed precision. In a standard survey sampling setting, λ^* minimizes $c \times \text{tr}(\text{V}_1/\text{N}) = (\text{Nc}_0 + \text{nc}_1) \times \text{tr}(\text{A}^{-1}/\text{N} + \text{A}^{-1}\text{BA}^{-1}/\text{n}) \propto \lambda^{-1}\text{tr}(\text{c}_1\text{A}^{-1}) + \lambda\text{tr}(\text{c}_0\text{A}^{-1}\text{BA}^{-1})$, which leads to $\lambda^* = [c_1\text{tr}(\text{A}^{-1})/\{\text{c}_0\text{tr}(\text{A}^{-1}\text{BA}^{-1})\}]^{1/2}$.

In Appendix, we show that $\widehat{\beta}_{\text{II}}$, the solution to (3.6), is also asymptotically unbiased and normally distributed as described in the following theorem.

Theorem 2 *Under the conditions listed in Appendix, we have*

$$(N+n)^{1/2}(\widehat{\beta}_{\text{II}} - \beta) \xrightarrow{D} N(0, V_2),$$

where $V_2 = \Gamma_2^{-1}\Sigma_2\Gamma_2^{-1}$, $\Gamma_2 = (\lambda A + C)/(1+\lambda)$, $\Sigma_2 = \Gamma_2 + \lambda^2 B/(1+\lambda)$, A *and* B *are defined in Theorem 1, and* $C = \text{Var}\{U(\beta; X, Y)\}$.

Remark 4 The asymptotic covariance matrix V_2 can be expressed as

$$V_2 = \Gamma_2^{-1} + \lambda^2\Gamma_2^{-1}B\Gamma_2^{-1}/(1+\lambda).$$

Here, Γ_2 is the weighted average of the Fisher information from the incomplete primary data assuming $f(x \mid w)$ is known and the Fisher information from the complete validation

data. When N is much larger than n, Γ_2 is close to Γ_1 defined in Theorem 1. The second part of V_2 is the extra price we pay for modeling $f(x \mid w)$ nonparametrically.

Remark 5 A plug-in estimator for V_2 can be derived similarly as Scenario I: we can estimate A and B using the method-of-moment estimators described in Remark 2 and estimate C by $\widehat{C} = n^{-1}\sum_{i=N+1}^{N+n} U^{\otimes 2}(\widehat{\beta}_{\text{II}}; X_i, Y_i)$. The optimal design of the study can be determined by minimizing the product of precision and cost if pilot estimates of A, B, and C are available, as discussed in Remark 3. However, under Scenario II, the optimal sample ratio, λ^*, needs to be solved numerically because it cannot be expressed in a closed form.

3.2.3 Extension to include an error-free covariate

In many applications, there is also an error-free predictor Z, which may be related to both Y and X. In this case, the independent variable is (X, Z) of dimension $p = p_1 + p_2$, where p_1 and p_2 are the dimensions of X and Z respectively. We assume that Z is available in both the primary and validation data set. Our methods are still applicable with a minor modification. Specifically, $\bar{U}_i(\beta)$ in estimating equations (3.5) and (3.6) is replaced by

$$\bar{U}_i(\beta) = \frac{\sum_{j=N+1}^{N+n} U(\beta; X_j, Z_i, Y_i) f(Y_i \mid X_j, Z_i; \beta) K_h\{(W_i - W_j), (Z_i - Z_j)\}}{\sum_{j=N+1}^{N+n} f(Y_i \mid X_j, Z_i; \beta) K_h\{(W_i - W_j), (Z_i - Z_j)\}}, \tag{3.9}$$

where $K(\cdot)$ is a p-variate kernel function of order k, and $U(\beta; x, z, y)$ is an unbiased estimating function for β such that $E\{U(\beta; X, Z, Y)\} = 0$. In (3.6), $U(\beta; X_i, Y_i)$ should also be replaced by $U(\beta; X_i, Z_i, Y_i)$. In the special case that Z is conditionally independent with X given W, $\bar{U}_i(\beta)$ can be simplified as

$$\bar{U}_i(\beta) = \frac{\sum_{j=N+1}^{N+n} U(\beta; X_j, Z_i, Y_i) f(Y_i \mid X_j, Z_i; \beta) K_h\{(W_i - W_j)\}}{\sum_{j=N+1}^{N+n} f(Y_i \mid X_j, Z_i; \beta) K_h\{(W_i - W_j)\}}, \tag{3.10}$$

where $K(\cdot)$ is a p_1-variate kernel function of order k.

In real applications, X is usually of low dimension, whereas Z can be of high dimension. Estimators based on (3.9) could suffer from the curse of dimensionality. When Z is conditionally independent with X given W, the curse of dimensionality can be circumvented by using (3.10). More discussions on possible solutions facing high dimensional covariate Z are provided in Section 3.6.

3.3 Computation and implementation issues

3.3.1 Trimming bound and bandwidth selection

We elaborate the implementation of our methods in the context of solving (3.5) under Scenario I. The same principle applies when solving (3.6) or when an error-free covariate Z is present.

In practice, N is usually much larger than n, so some W_i in the denominator of (3.4) may step outside the hull of $\{W_j\}_{j=N+1}^{N+n}$ in the validation data, which results in a zero or near-zero denominator in (3.4). To avoid numerical instability, we replace the denominator of (3.4) by $\max\{\sum_{j=N+1}^{N+n} f(Y_i \mid X_j;\beta)K_h(W_i - W_j), nt_n\}$ where $t_n > 0$ is a trimming bound. We find that $t_n = n^{-5}$ leads to satisfactory results in our simulation studies and data analysis. Sepanski and Lee (1995) fix the trimming bound at a small number such as 10^{-5}, whereas Wang and Yu (2007) let the trimming bound vary with n as we do. We have tried different trimming bounds in our numerical studies and the results are almost identical as long as t_n is small enough.

Another important issue in implementing the proposed methods is bandwidth selection. We propose to set the bandwidth using the empirical formula $h = \widehat{\sigma}_W n^{-1/(p_1+k+1)}$, where $\widehat{\sigma}_W$ is the estimated standard deviation of W in the validation data set, $p_1 = 1$, and k is the order of the kernel function. Carroll and Wand (1991) also use a similar empirical formula to set the bandwidth when X is a univariate predictor and $K(\cdot)$ is a second order kernel function. When $p_1 > 1$, different smoothing parameters can be

used in different components. For example, for the j-th component, one can use the bandwidth $h_j = \widehat{\sigma}_W^{[j]} n^{-1/(p_1+k+1)}$, where $\widehat{\sigma}_W^{[j]}$ is the estimated standard deviation of the j-th component of W. When an error-free covariate Z exists, in the special case that Z is conditionally independent with X given W, the empirical formula for the bandwidth does not change. Otherwise, similar empirical formula can be defined by replacing p_1 by p (e.g. Section 3.4.2). The proposed empirical bandwidth satisfies the conditions listed in Appendix. Our simulation studies show that the numerical performance of the proposed methods is not very sensitive to the choice of bandwidth and the proposed empirical formula works well.

3.3.2 Algorithm and connection with fractional imputation

Our methods in Section 3.2.1 are based on an approximation of the expected estimating function (3.3). We can express the approximate expected estimating function $\bar{U}_i(\beta)$ as

$$\bar{U}_i(\beta) = \widehat{E}\{U(\beta; X_i, Y_i) \mid Y_i, W_i; \beta\}$$

and solve (3.5) using an EM type of iterative procedure. Let $\widehat{\beta}_{\mathrm{I}}^{(t)}$ be the value of $\widehat{\beta}_{\mathrm{I}}$ at the t-th iteration, then let $\widehat{\beta}_{\mathrm{I}}^{(t+1)}$ be the solution of

$$\textstyle\sum_{i=1}^{N} \widehat{E}\{U(\beta; X_i, Y_i) \mid Y_i, W_i; \widehat{\beta}_{\mathrm{I}}^{(t)}\} = 0.$$

Equivalently, the equation above is expressed as

$$\textstyle\sum_{i=1}^{N} \sum_{j=1}^{n} w_{ij}^{*(t)} U(\beta; X_i^{*(j)}, Y_i) = 0, \tag{3.11}$$

where $X_i^{*(j)} = X_{N+j}$ and

$$w_{ij}^{*(t)} = \frac{f(Y_i \mid X_i^{*(j)}; \widehat{\beta}_{\mathrm{I}}^{(t)}) K_h(W_i - W_{N+j})}{\sum_{k=1}^{n} f(Y_i \mid X_i^{*(k)}; \widehat{\beta}_{\mathrm{I}}^{(t)}) K_h(W_i - W_{N+k})}, \tag{3.12}$$

for $j = 1, \ldots, n$. For each X_i, we are essentially creating n imputed values $X_i^{*(j)}$ with weights $w_{ij}^{*(t)}$, $j = 1, \cdots, n$. By choosing $X_i^{*(j)} = X_{N+j}$, equation (3.11) can be regarded

as a semiparametric version of fractional imputation (Kim, 2011; Kim and Shao, 2013). The weight in (3.12) is called the fractional weight and it reflects the point mass assigned to the imputed value $X_i^{*(j)}$. The fractional weights satisfy $\sum_{j=1}^{n} w_{ij}^{*(t)} = 1$ for all $i = 1, \cdots, N$.

Given the value of $\widehat{\beta}_{\mathrm{I}}^{(t)}$, we can solve the imputed equation (3.11) using the usual Newton-Raphson method and iterate until the value of $\widehat{\beta}_{\mathrm{I}}$ converges. Similar algorithms can be developed for solving the estimating equation (3.6) and when an error-free covariate Z is observed as described in Section 3.2.3.

3.4 Simulation studies

3.4.1 Simulation 1

In our first simulation study, the model only includes a univariate error-prone predictor. We simulate Y from a conditional log-normal distribution $[\log(Y) \mid X] \sim \mathrm{N}(\beta_0 + \beta_1 X, \sigma^2)$, where $X \sim \mathrm{Unif}(0, 1)$, $(\beta_0, \beta_1) = (1, -1)$ and $\sigma = 1/4$. Let $U(\beta; x, y)$ be the score function $S(\beta; x, y) = \{\log(y) - \beta_0 - \beta_1 x\}(1, x)^{\mathrm{T}}$. The surrogate W is generated from $N(5X/4, \sigma_e^2)$, where $\sigma_e = 3\sigma_X/4$ and σ_X is the standard deviation of X. The sample size for the primary data is fixed at $N = 500$. For the validation sample size, we consider $n = 50, 100$ and 250, corresponding to three sample size ratios $\lambda_N = N/n = 10, 5$ and 2. We use the Epanechnikov kernel, $K_0(x) \propto (1 - x^2) I\{|x| \leq 1\}$, which is a second order kernel function ($k = 2$). Denote the marginal variance of W as σ_W^2 and we set the bandwidth as $h^* = \widehat{\sigma}_W n^{-1/4}$ following the discussion in Section 3.3.1. For comparison, we also present the estimation results of the competitive estimators of Wang and Yu (2007), denoted by $\widehat{\beta}_{\mathrm{I}}^{\mathrm{WY}}$ and $\widehat{\beta}_{\mathrm{II}}^{\mathrm{WY}}$ for Scenarios I and II respectively.

Table 3.1 summarizes the Monte Carlo bias, standard deviation and root mean squared error (RMSE) for the two estimators under both Scenarios, based on 1,000 replications. The proposed estimators remarkably outperform those of Wang and Yu

(2007) in terms of RMSE in all scenarios. Compared with the estimators of Wang and Yu (2007), the proposed estimators achieve greater reduction of bias with very little cost of increase in standard deviation. The empirical biases of our estimators are much smaller than the standard deviations, which confirms our theory that our estimators are asymptotically unbiased. In contrast, the biases in the estimators of Wang and Yu (2007) are distinctively large and dominate their RMSE's. To better illustrate these results using graphs, we show box plots of the estimates of β_1 in Figure 3.1. These plots clearly show that the estimators of Wang and Yu (2007) suffer from substantial biases, while the biases of our estimators are negligible.

We also perform two sensitivity analyses to our methods. We first investigate the sensitivity of our proposed methods to different choices of bandwidths. We repeat the estimation with $h = \rho \times h^*$ with different values of ρ, where $h^* = \widehat{\sigma}_W n^{-1/4}$ is the bandwidth using our empirical formula. In Figure 3.2, we plot the mean squared error of the proposed estimators under different scenarios and sample size ratios against ρ. As we can see, the estimators are optimal or near optimal at the proposed choice of bandwidth for $\rho = 1$, and they are not very sensitive to the bandwidth as long as h is not too small. This phenomenon is more prominent when λ_N is relatively large, which is usually the case in real data. Furthermore, compared with the estimates of β_0, the estimates of β_1 are slightly more sensitive because of the measurement errors.

Next, we investigate the robustness of our estimators against misspecification of the likelihood $f(y \mid x; \beta)$. The setting of the simulation study is similar as above except that Y is generated from $[\log(Y) \mid X] \sim \beta_0 + \beta_1 X + \sigma \mathrm{t}_\nu$, where t_ν is a t random variable with $\nu = 5$ degrees of freedom. We consider two versions of our proposed estimator: $\widehat{\beta}^{\mathrm{Mis}}$ is the estimator with $U(\beta; x, y)$ and $f(y \mid x; \beta)$ being the score and likelihood under log-normal assumptions, which is a misspecification under the current setting; $\widehat{\beta}$ is obtained when $U(\beta; x, y)$ and $f(y \mid x; \beta)$ are the score and likelihood under the correctly specified log-t model. Table 3.2 summarizes the results of the two estimators. Compared with the

Table 3.1 Results of Simulation 1 based on 1,000 replications

	$N = 500, n = 50$			$N = 500, n = 100$			$N = 500, n = 250$		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Scenario I									
$\widehat{\beta}_{0,\mathrm{I}}^{\mathrm{WY}}$	−129	35	134	−128	30	131	−125	28	128
$\widehat{\beta}_{0,\mathrm{I}}$	−8	44	44	−6	36	37	−2	32	32
$\widehat{\beta}_{1,\mathrm{I}}^{\mathrm{WY}}$	256	58	263	256	48	260	249	44	253
$\widehat{\beta}_{1,\mathrm{I}}$	13	75	76	12	61	62	3	52	52
Scenario II									
$\widehat{\beta}_{0,\mathrm{II}}^{\mathrm{WY}}$	−117	32	121	−107	26	110	−83	22	86
$\widehat{\beta}_{0,\mathrm{II}}$	−7	39	39	−5	31	31	−1	23	23
$\widehat{\beta}_{1,\mathrm{II}}^{\mathrm{WY}}$	232	54	238	213	43	218	166	35	169
$\widehat{\beta}_{1,\mathrm{II}}$	11	67	67	11	53	54	2	38	38

Note: $(\widehat{\beta}_{0,\mathrm{I}}^{\mathrm{WY}}, \widehat{\beta}_{1,\mathrm{I}}^{\mathrm{WY}})$ and $(\widehat{\beta}_{0,\mathrm{II}}^{\mathrm{WY}}, \widehat{\beta}_{1,\mathrm{II}}^{\mathrm{WY}})$ are the competitive estimators of Wang and Yu (2007) for (β_0, β_1) under Scenario I and Scenario II. Bias, the empirical bias ($\times 10^3$); SD, the empirical standard deviation ($\times 10^3$); RMSE, the empirical root mean square error ($\times 10^3$).

estimators under correctly specified model, the estimators under misspecification only show less than 10% of increases in RMSE, which also shows that our estimator is robust.

3.4.2 Simulation 2

We conduct a second simulation study with bivariate predictors ($p = 2$), which is in line with the setting of Section 3.2.3 and our real data application in Section 3.5. The data are generated from a log-linear model $[\log(Y) \mid X, Z] \sim \mathrm{N}(\beta_1 X + \beta_2 Z, \sigma^2)$, where $(\beta_1, \beta_2) = (-1, 1)$ and $\sigma = 1/4$. The error-prone covariate X and its surrogate W are simulated in the same way as in Section 3.4.1. We let Z be error-free and consider two situations where Z is either independent or dependent with X. Even though Wang and Yu (2007) never considered an error-free covariate Z, we extend their methods to this setting for a comparison.

First, we consider a relatively simple case where Z is generated independently from a normal distribution $\mathrm{N}(\mu_Z, \sigma_Z^2)$, with $\mu_Z = \mu_X$ and $\sigma_Z^2 = \sigma_X^2$. We perform estimation

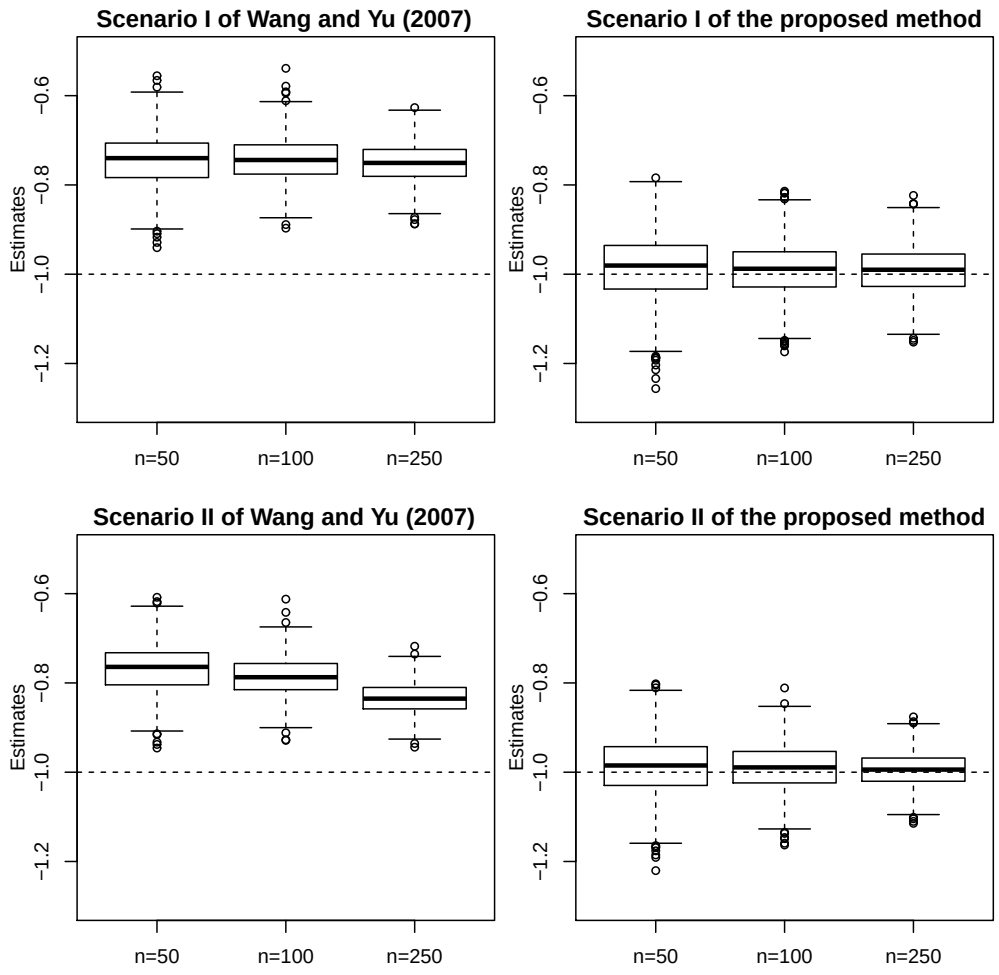


Figure 3.1 Box plots for the estimates of β_1 using the method of Wang and Yu (2007) and the proposed method. The dashed horizontal line denotes the true value of β_1.

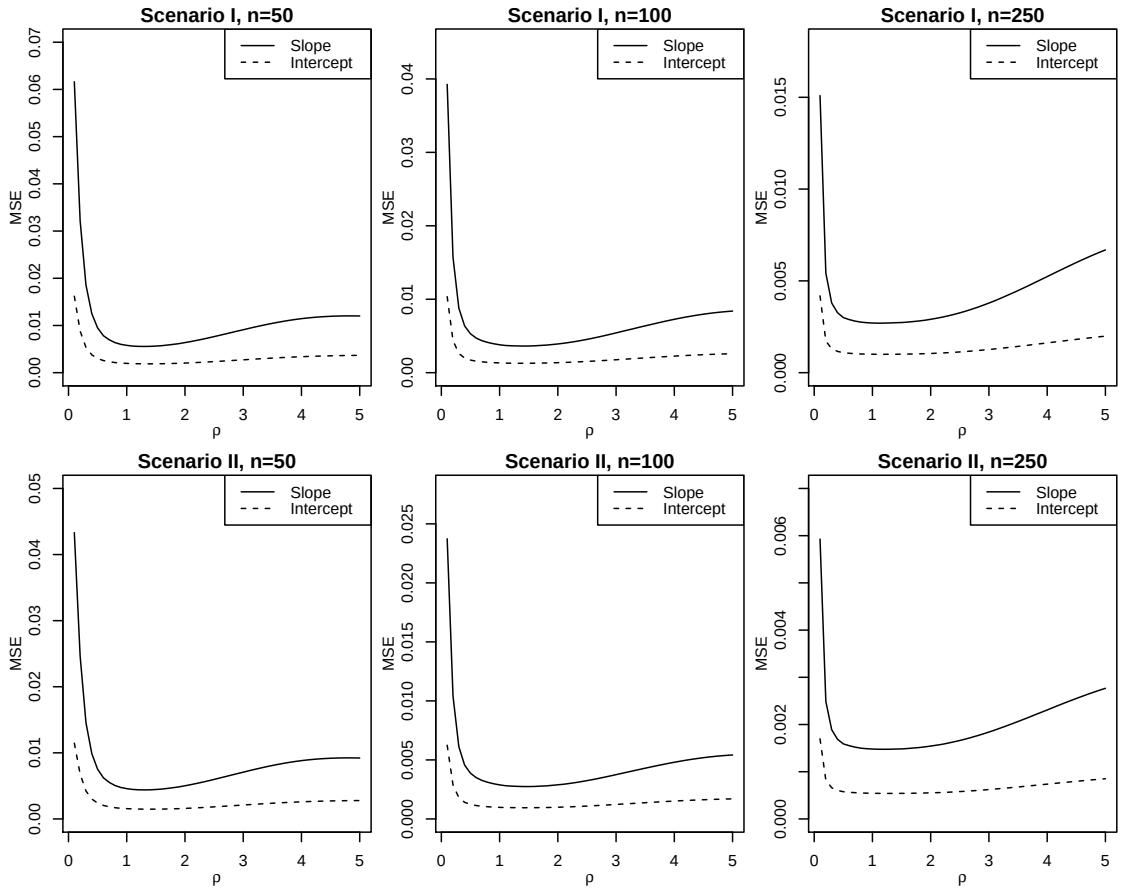


Figure 3.2 Sensitivity analysis for the mean square error (MSE) of the estimates for β_0 (dashed curves) and β_1 (solid curves) using the proposed method against different bandwidths. For each sub-figure, $\rho = 1$ means the proposed bandwidth h^* is in use.

Table 3.2 Sensitivity analysis results based on 1,000 replications

	$N = 500, n = 50$			$N = 500, n = 100$			$N = 500, n = 250$		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Scenario I									
$\widehat{\beta}^{\text{Mis}}_{0,\text{I}}$	−16	49	52	−13	42	44	−8	36	37
$\widehat{\beta}_{0,\text{I}}$	−8	48	49	−5	40	41	−1	34	34
$\widehat{\beta}^{\text{Mis}}_{1,\text{I}}$	32	85	90	22	71	74	15	62	64
$\widehat{\beta}_{1,\text{I}}$	18	83	85	8	68	68	4	58	58
Scenario II									
$\widehat{\beta}^{\text{Mis}}_{0,\text{II}}$	−14	45	47	−10	36	37	−4	27	28
$\widehat{\beta}_{0,\text{II}}$	−7	44	44	−4	34	34	0	25	25
$\widehat{\beta}^{\text{Mis}}_{1,\text{II}}$	28	77	82	17	61	63	8	47	47
$\widehat{\beta}_{1,\text{II}}$	15	75	77	5	58	58	0	42	42

Note: The layout of this table is similar to that of Table 3.1. $\widehat{\beta}^{\text{Mis}}$ and $\widehat{\beta}$ are different versions of our estimator under either misspecified or correctly specified regression model.

using estimating equations based on (3.10), the Epanechnikov kernel and the same bandwidth as described in Section 3.4.1. The results based on 1,000 simulation replicates are reported in Table 3.3, where we summarize the bias, standard deviation and RMSE for both estimators under comparsion. These results reveal a similar pattern as those in Table 3.1: the estimators of Wang and Yu (2007) suffer from substantial biases; our estimators show much smaller biases with comparable standard deviations and perform favorably in terms of RMSE. We have also examined the box plots for the estimators and performed a sensitivity analysis. The results are similar to those in Figure 3.1, Figure 3.2, and Table 3.2, so they are omitted for conciseness.

We then consider a more complex yet more realistic case where Z is dependent with X. Specifically, we generate $Z = -0.2X + 1.2U$, where $U \sim \text{N}(\mu_U, \sigma_U^2)$ with $\mu_U = \mu_X$ and $\sigma_U^2 = \sigma_X^2$. In this setting, Z and X have a negative correlation of -0.16, which is similar to the real data in Section 3.5. Our estimators are based on (3.9) with a bivariate kernel function. We choose a product Epanechnikov kernel $K(w, z) = K_0(w)K_0(z)$

Table 3.3 Results for Simulation 2 when Z is independent with X

	$N = 500, n = 50$			$N = 500, n = 100$			$N = 500, n = 250$		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Scenario I									
$\widehat{\beta}_{1,\mathrm{I}}^{\mathrm{WY}}$	146	47	153	144	39	149	142	35	146
$\widehat{\beta}_{1,\mathrm{I}}$	8	51	52	5	44	44	1	37	37
$\widehat{\beta}_{2,\mathrm{I}}^{\mathrm{WY}}$	−108	44	117	−109	40	116	−108	36	113
$\widehat{\beta}_{2,\mathrm{I}}$	−5	45	45	−5	41	42	−2	37	37
Scenario II									
$\widehat{\beta}_{1,\mathrm{II}}^{\mathrm{WY}}$	132	43	138	120	34	125	94	27	98
$\widehat{\beta}_{1,\mathrm{II}}$	7	47	47	4	37	37	1	29	29
$\widehat{\beta}_{2,\mathrm{II}}^{\mathrm{WY}}$	−98	41	106	−90	34	97	−71	27	76
$\widehat{\beta}_{2,\mathrm{II}}$	−4	42	42	−4	35	35	−1	28	28

Note: Same layout as Table 3.1 and the results are based on 1,000 replications.

with different bandwidths in w and z coordinates. We set $h^W = \widehat{\sigma}_W n^{-1/5}$ for W and $h^Z = \widehat{\sigma}_Z n^{-1/5}$ for Z. The results from 1,000 simulation replicates are summarized in Table 3.4. As we can see, the proposed estimators still perform much better than the competitive estimators in terms of both bias and RMSE.

3.5 Data analysis

We now illustrate the proposed method by applying it to a data set from the Korean Longitudinal Study of Aging (KLoSA). Since the proportion of elderly citizens has been increasing rapidly in South Korea, from 5.9% in 1995 to 8.7% in 2004, health issues related to the elderly have received more and more attention in recent years. The KLoSA is a longitudinal survey conducted by Korean Labor Institute every two years starting from 2006 for the citizens in South Korean aged 45 or over. Details about the study can be found at *http://www.kli.re.kr/klosa/en/about/introduce.jsp*. The data considered in this paper are based on the survey conducted in 2006, including a primary data set ($N = 9842$) and a validation data set ($n = 505$). The validation data set is a random

Table 3.4 Results for Simulation 2 when Z is dependent with X

	$N = 500, n = 50$			$N = 500, n = 100$			$N = 500, n = 250$		
Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Scenario I									
$\widehat{\beta}_{1,\mathrm{I}}^{\mathrm{WY}}$	131	56	143	125	43	132	118	35	123
$\widehat{\beta}_{1,\mathrm{I}}$	52	61	80	36	47	60	22	39	44
$\widehat{\beta}_{2,\mathrm{I}}^{\mathrm{WY}}$	−90	56	106	−85	44	95	−79	37	87
$\widehat{\beta}_{2,\mathrm{I}}$	−35	60	69	−25	47	53	−16	39	42
Scenario II									
$\widehat{\beta}_{1,\mathrm{II}}^{\mathrm{WY}}$	116	51	126	101	36	107	75	27	80
$\widehat{\beta}_{1,\mathrm{II}}$	45	55	71	29	39	48	12	28	31
$\widehat{\beta}_{2,\mathrm{II}}^{\mathrm{WY}}$	−78	51	93	−67	36	76	−50	26	58
$\widehat{\beta}_{2,\mathrm{II}}$	−30	53	61	−19	38	43	−9	27	28

Note: Same layout as Table 3.1 and the results are based on 1,000 replications.

sample of the whole data set with $N + n = 10347$ individuals. The demographic and health condition information of the individuals enrolled in the study has been collected.

One of the most important objectives of the study is to investigate risk factors for common geriatric diseases. In our analysis, we investigate the relationship between hypertension and two risk factors, age and body mass index (BMI), where BMI is defined as one's weight divided by the square of one's height. The response Y is an indicator of hypertension status with $Y = 1$ meaning hypertension, and the error-free predictor Z is the age. The primary data set includes $N = 9842$ subjects, where the true BMI X is not available but instead an approximate BMI W can be computed from the self-reported weight and height for every individual. The study also has a Type II validation data set, consisting of (Y, X, W, Z) from $n = 505$ subjects where the true BMI X is available because there exist precise physical measurements on the height and weight for every individual. The sample size ratio of the primary data and validation data is $\lambda_N = 18.78$.

To gain further insight into the relationship between X and W, we generate some summary graphs based on the validation data, as shown in Figure 3.3. The left panel

of Figure 3.3 shows a linear regression of W against X compared with the reference line $w = x$. Interestingly, the graph suggests that most people tend to under-report their BMI and some people with low BMI tend to over-report their values. In the right panel of Figure 3.3, the normal Q-Q plot for the residuals from the regression analysis implies that an additive normal measurement error model is not appropriate (the Shapiro-Wilk test of normality shows a p-value of 4.60×10^{-9}). These results stress the importance of semiparametric methods that do not rely on structural or distributional assumptions on the measurement errors. Our proposed methods have such merit of robustness.

We consider a logistic regression model, $\text{logit}\{P(Y = 1 \mid X, Z)\} = \beta_0 + \beta_1 X + \beta_2 Z$, where $\text{logit}(t) = \log\{t/(1-t)\}$. The correlation coefficient between X and Z based on the validation data is -0.11, which suggests that the relationship between X and Z is small or inclusive. We apply two version of the proposed method with or without the assumption of conditional independence between X and Z, and the two versions are based on (3.10) and (3.9) respectively. In either case, we set the trimming bound and bandwidth as described in Section 3.3. For comparison, we also consider the naive estimator using (W, Z) as the predictor and the estimator of Wang and Yu (2007). For Wang and Yu's method, we also consider two versions with or without the assumption of conditional independence between X and Z. For a fair comparison, standard errors for all methods are estimated using bootstrapping. Specifically, we adopt a nonparametric bootstrap method (Efron and Tibshirani, 1993) by sampling N subjects with replacement from the primary data set and n subjects with replacement from the validation data set. For each method under consideration, we apply the estimator to 1,000 bootstrap samples and estimated the standard error by the standard deviation of the 1,000 estimates.

The data analysis results using different methods are summarized in Table 3.5. All methods show similar results on β_2, indicating age has a significant positive effect on the risk of hypertension. These results are consistent across different methods because age is error free and has little correlation with BMI. In contrast, the naive estimator for β_1 is

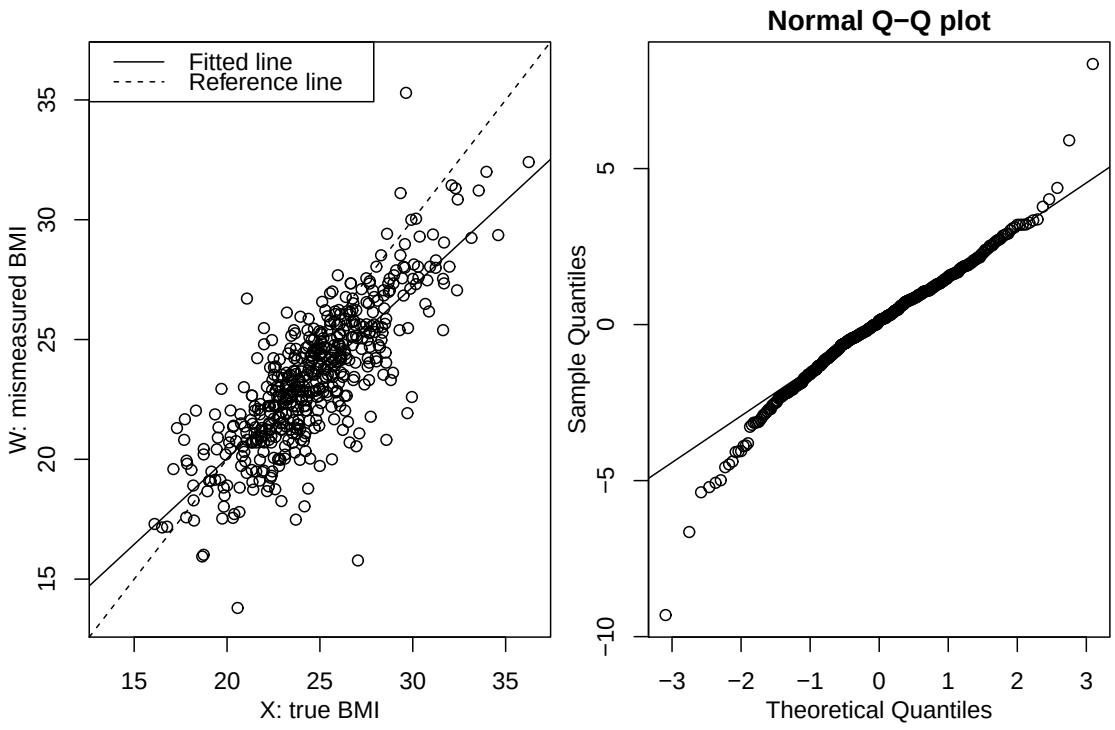


Figure 3.3 Preliminary analysis for the measurement errors using the validation data. The left panel is a scatter plot of W (estimated BMI based on self-reported information) against X (true BMI). The solid line represents the fitted regression line, $w = 5.70 + 0.72x$, and the dashed line is the reference line, $w = x$. The right panel is the normal Q-Q plot for the residuals obtained from the simple linear regression.

significantly attenuated toward 0. The attenuation bias is so severe that a test based on the naive estimator would declare BMI as an insignificant predictor. Such attenuation effect of measurement errors is well-documented in the literature and hence expected. Both the methods of Wang and Yu (2007) and ours offer some degree of bias correction and declare BMI has a significant positive association with hypertension, which means increase in BMI leads to higher disease risk. Similar to what we observe in our simulation studies, our methods offer more bias correction than those of Wang and Yu (2007). Since the correlation between X and Z is weak in this data set, methods with or without the conditional independence assumption do not show much difference in the results.

Table 3.5 Data analysis results for the KLoSA data using various methods

	β_0		β_1		β_2	
Method	Estimates	SE	Estimates	SE	Estimates	SE
Naive	−4.8068	0.4042	0.0205	0.1441	0.0534	0.0022
Wang and Yu	−7.5132	0.2503	0.1203	0.0074	0.0571	0.0022
Wang and Yu*	−7.8978	0.2976	0.1210	0.0087	0.0632	0.0029
Proposed	−9.3703	0.3605	0.1844	0.0116	0.0612	0.0023
Proposed*	−9.4157	0.4057	0.1734	0.0126	0.0667	0.0031

Note: Naive, the naive method using (W, Z) as the predictor; Wang and Yu and Wang and Yu*, the method of Wang and Yu (2007) with or without the assumption of conditional independence between X and Z; Proposed and Proposed*, the proposed method with or without the assumption of conditional independence between X and Z; SE, bootstrap standard error based on 1,000 bootstrap samples.

3.6 Concluding remarks

Since our proposed methodology is based on kernel smoothing, it might suffer from the curse of dimensionality when the dimension of the predictors p is high. A common situation is that the dimension of X is low but the dimension of Z is high. For example, X is one's long-term nutrition intake level and Z includes genetic information. Usually, Z can be decomposed into two parts, $Z = (Z_1^{\mathrm{T}}, Z_2^{\mathrm{T}})^{\mathrm{T}}$, Z_1 is related to X and is of low dimension, and Z_2 is independent of X given W. Using similar arguments as in Section 3.2.3, we only need to include Z_1 in the kernel weight and thus avoid the curse of dimensionality. To choose Z_1, we can apply variable selection methods, such as Lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), or sure independence screening (Fan and Lv, 2008), to the validation data. Another possibility to circumvent the curse of dimensionality is to apply the additive model (Hastie and Tibshirani, 1990) to the conditional expectation in (3.3). Further more, a hybrid method of the additive model and variable selection, such as that proposed by Huang, Horowitz, and Wei (2010), can also be applied to the validation data.

In this article, the measurement error model is modeled nonparametrically. When a classical measurement error model is appropriate, likelihood-based inference described in Carroll et al. (2006) can be made. When repeated measurements are available, nonparametric estimation of the measurement error model using a deconvolution method (Delaigle, Hall, and Meister, 2008) can also be employed.

3.7 Acknowledgments

Kim's research is supported in part by a grant from U.S. National Science Foundation, MMS-121339. Li's research is partially supported by the U.S. National Science Foundation, award DMS-1317118.

3.8 Appendix: technical details

Regularity conditions

(C1) The kernel function $K(\cdot)$ is bounded and symmetric about zero over a bounded support and of order $k(\geq \max(2, p_1))$.

(C2) The joint density $f(x, w)$ is $k+1$ times continuously differentiable with respect to w.

(C3) The bandwidth h satisfies: $nh^{2p_1} \to \infty$ and $nh^{2(k+1)} \to 0$.

(C4) $\lambda_N = N/n \to \lambda(< \infty)$.

Conditions (C1) and (C2) are standard regularity conditions in nonparametric regression. (C3) is the bandwidth condition used in Theorem 1 and Theorem 2. Compared with Carroll and Wand (1991) and Wang and Yu (2007), our bandwidth condition is refined so that lower order kernel functions can be used. (C4) implies that the information from the primary data set and the information from the validation data set is comparable.

Proof of Theorem 1

For the estimating equation (3.5), we apply Taylor's theorem to get

$$\widehat{\beta}_{\text{I}} - \beta = J_{nN}^{-1}(\beta^*)\bar{U}_{nN}(\beta), \tag{3.13}$$

where $J_{nN}(\beta^*) = N^{-1}\sum_{i=1}^{N} -\partial/\partial\beta^{\text{T}}\{\bar{U}_i(\beta)\} \mid_{\beta=\beta^*}$ with $\| \beta^* - \beta \| \leq \| \widehat{\beta}_{\text{I}} - \beta \|$, and $\bar{U}_{nN}(\beta) = N^{-1}\sum_{i=1}^{N}\bar{U}_i(\beta)$.

Define $\mu_c(\beta; y, w) = E\{\bar{c}(\beta; y, w)\}$ and $\mu_d(\beta; y, w) = E\{\bar{d}(\beta; y, w)\}$. It follows from standard arguments of multivariate kernel estimation (Fan and Gijbels, 1996) that

$$\mu_c(\beta; y, w) = c(\beta; y, w) + O(h^k), \quad \mu_d(\beta; y, w) = d(\beta; y, w) + O(h^k). \tag{3.14}$$

By the definition of $\bar{U}_{nN}(\beta)$,

$$N^{1/2}\bar{U}_{nN}(\beta) = N^{-1/2}\textstyle\sum_{i=1}^{N}\bar{e}(\beta; Y_i, W_i) = N^{-1/2}\sum_{i=1}^{N}\bar{d}(\beta; Y_i, W_i)/\bar{c}(\beta; Y_i, W_i). \tag{3.15}$$

Suppressing the inner part of $(\beta; Y_i, W_i)$, we have

$$\begin{aligned} \frac{\bar{d}}{\bar{c}} &= \frac{\mu_d + (\bar{d} - \mu_d)}{\mu_c + (\bar{c} - \mu_c)} \\ &= \frac{\mu_d}{\mu_c} + \frac{\bar{d} - \mu_d}{\mu_c} - \frac{\mu_d}{\mu_c}\frac{\bar{c} - \mu_c}{\mu_c} + h.o.t \\ &:= R_1 + R_2 - R_3 + R_4. \end{aligned} \tag{3.16}$$

First, we proceed to show that

$$N^{-1/2}\textstyle\sum_{i=1}^{N}R_1(\beta; Y_i, W_i) = N^{-1/2}\sum_{i=1}^{N}e(\beta; Y_i, W_i) + O_p(h^k + N^{1/2}h^{k+1}). \tag{3.17}$$

For simplicity, we only show this result for $p_1 = 1$ but the conclusion holds for arbitrary p_1 by similar arguments. In detail, $\mu_c(\beta; y, w)$ and $\mu_d(\beta; y, w)$ in (3.14) can be expressed as

$$\begin{aligned} \mu_c(\beta; y, w) &= c(\beta; y, w) + \frac{\sigma_k h^k}{k!}q_c(\beta; y, w) + O_p(h^{k+1}), \\ \mu_d(\beta; y, w) &= d(\beta; y, w) + \frac{\sigma_k h^k}{k!}q_d(\beta; y, w) + O_p(h^{k+1}), \end{aligned}$$

where

$$\begin{aligned}\sigma_k &= \int K(u)u^k \, \mathrm{d}u, \\ q_c(\beta; y, w) &= \int f(y \mid x; \beta) f^{(0,k)}(x, w) \, \mathrm{d}x, \\ q_d(\beta; y, w) &= \int U(\beta; x, y) f(y \mid x; \beta) f^{(0,k)}(x, w) \, \mathrm{d}x,\end{aligned}$$

and $f^{(0,k)}(x, w)$ is the k-th order partial derivative of $f(x, w)$ with respect to w. Then, we obtain

$$\frac{\mu_d(\beta; y, w)}{\mu_c(\beta; y, w)} = \frac{d(\beta; y, w)}{c(\beta; y, w)} - \frac{\sigma_k h^k}{k!} \left\{ \frac{q_d(\beta; y, w)c(\beta; y, w) - q_c(\beta; y, w)d(\beta; y, w)}{c^2(\beta; y, w)} \right\} + O_p(h^{k+1}).$$

Since it can be verified that

$$E\{q_d(\beta; Y, W)c(\beta; Y, W)/c^2(\beta; Y, W)\} = E\{q_c(\beta; Y, W)d(\beta; Y, W)/c^2(\beta; Y, W)\},$$

by the central limit theorem, (3.17) can be proved.

By (3.14), (C3), and the central limit theorem, it can be shown that

$$\begin{aligned}N^{-1/2}\textstyle\sum_{i=1}^N R_2(\beta; Y_i, W_i) &= N^{-1/2}\textstyle\sum_{i=1}^N \dfrac{\bar{d}(\beta; Y_i, W_i) - \mu_d(\beta; Y_i, W_i)}{c(\beta; Y_i, W_i)} + O_p(h^{k-p_1/2}), \\ N^{-1/2}\textstyle\sum_{i=1}^N R_3(\beta; Y_i, W_i) &= N^{-1/2}\textstyle\sum_{i=1}^N \dfrac{d(\beta; Y_i, W_i)\{\bar{c}(\beta; Y_i, W_i) - \mu_c(\beta; Y_i, W_i)\}}{c^2(\beta; Y_i, W_i)} \\ &\quad + O_p(h^{k-p_1/2}), \\ N^{-1/2}\textstyle\sum_{i=1}^N R_4(\beta; Y_i, W_i) &= N^{1/2} O_p\{(nh^{p_1})^{-1}\} = o_p(1).\end{aligned}$$

Therefore, (3.16) implies that (3.15) can be written as a two sample statistic

$$N^{1/2}\bar{S}_{nN}(\beta) = n^{-1}N^{-1/2}\textstyle\sum_{i=1}^N \sum_{j=N+1}^{N+n} \psi_n(\beta; U_i, V_j) + o_p(1), \tag{3.18}$$

where $U_i = (Y_i, W_i)$, $V_j = (X_j, W_j)$, and

$$\begin{aligned}\psi_n(\beta; U_i, V_j) &= e(\beta; Y_i, W_i) + \frac{U(\beta; X_j, Y_i) f(Y_i \mid X_j; \beta) K_h(W_i - W_j) - \mu_d(\beta; Y_i, W_i)}{c(\beta; Y_i, W_i)} \\ &\quad - \frac{d(\beta; Y_i, W_i)}{c^2(\beta; Y_i, W_i)} \left\{ f(Y_i \mid X_j; \beta) K_h(W_i - W_j) - \mu_c(\beta; Y_i, W_i) \right\}.\end{aligned}$$

By interchanging expectation and differentiation, it is not hard to see

$$E\{\psi_n(\beta; U_i, V_j) \mid U_i\} = e(\beta; Y_i, W_i). \tag{3.19}$$

Then, we have

$$E\{\psi_n(\beta; U_i, V_j)\} = E\{e(\beta; Y_i, W_i)\} = E\{U(\beta; X_i, Y_i)\} = 0. \tag{3.20}$$

Recall that $M_1(\beta; x) = E\{U(\beta; X, Y)|X = x\}$, $M_2(\beta; x, w) = E\{e(\beta; Y, W)|X = x, W = w\}$. By (3.14) and the definition of $\psi_n(\beta; U_i, V_j)$, we have

$$\begin{aligned}
& E\left\{\frac{U(\beta; X_j, Y_i) f(Y_i \mid X_j; \beta) K_h(W_i - W_j) - \mu_d(\beta; Y_i, W_i)}{c(\beta; Y_i, W_i)} \mid V_j\right\} \\
= \; & E\left\{U(\beta; X_j, Y_i) f(Y_i \mid X_j; \beta) K_h(W_i - W_j)/c(\beta; Y_i, W_i) \mid V_j\right\} + O_p(h^k) \\
= \; & \int\int \left\{U(\beta; X_j, y) f(y \mid X_j; \beta) K_h(w - W_j)\right\} \, \mathrm{d}w\mathrm{d}y + O_p(h^k) \\
= \; & M_1(\beta; X_j) + O_p(h^k).
\end{aligned}$$

Similarly, it can be verified that

$$\begin{aligned}
& E\left[\frac{d(\beta; Y_i, W_i)}{c^2(\beta; Y_i, W_i)} \left\{f(Y_i \mid X_j; \beta) K_h(W_i - W_j) - \mu_c(\beta; Y_i, W_i)\right\} \mid V_j\right] \\
= \; & M_2(\beta; X_j, W_j) + O_p(h^k).
\end{aligned}$$

Therefore, it follows that

$$E\{\psi_n(\beta; U_i, V_j) \mid V_i\} = M(\beta; X_j, W_j) + O_p(h^k). \tag{3.21}$$

Combining (3.19), (3.20), (3.21) and using Theorem B.1 of Sepanski and Lee (1995), we obtain

$$N^{1/2} \bar{U}_{nN}(\beta) \xrightarrow{D} N(0, \Sigma_1), \tag{3.22}$$

where $\Sigma_1 = \mathrm{Var}\{e(\beta; Y, W)\} + \lambda \mathrm{Var}\{M(\beta; X, W)\}$. Equations (3.18) and (3.20) indicate that the estimating equation (3.5) is asymptotically unbiased. Under regularity conditions for Z-estimators (van der Vaart, 1998), estimating equation (3.5) yields a sequence of solutions $\widehat{\beta}_{\mathrm{I}}$ which converges in probability to β. By (3.16) and (C3), we obtain

$$\bar{e}(\beta; Y_i, W_i) = e(\beta; Y_i, W_i) + O_p\{(nh^{p_1})^{-1/2} + h^k\} = e(\beta; Y_i, W_i) + o_p(1). \tag{3.23}$$

By (3.23), the consistency of $\widehat{\beta}_{\text{I}}$, and some smoothness conditions of $J_{nN}(\cdot)$, we have

$$J_{nN}(\beta^*) \xrightarrow{p} \Gamma_1, \tag{3.24}$$

where $\Gamma_1 = E[\partial/\partial\beta\{e(\beta; Y, W)\}] = \text{Var}\{e(\beta; Y, W)\}$. Theorem 1 follows from (3.13), (3.22), (3.24) and Slutsky's theorem.

Proof of Theorem 2

For the estimating equation (3.6), we use Taylor's theorem to obtain

$$\widehat{\beta}_{\text{II}} - \beta = \widetilde{J}_{nN}^{-1}(\beta^*)\widetilde{U}_{nN}(\beta),$$

where $|| \beta^* - \beta || \leq || \widehat{\beta}_{\text{II}} - \beta ||$, and

$$\begin{aligned} \widetilde{J}_{nN}(\beta) &= -(N+n)^{-1}\partial/\partial\beta^{\text{T}} \left\{ \textstyle\sum_{i=1}^{N} \bar{U}_i(\beta) + \sum_{i=N+1}^{N+n} \{U(\beta; X_i, Y_i)\} \right\}, \\ \widetilde{U}_{nN}(\beta) &= (N+n)^{-1} \left\{ \textstyle\sum_{i=1}^{N} \bar{U}_i(\beta) + \sum_{i=N+1}^{N+n} U(\beta; X_i, Y_i) \right\}. \end{aligned}$$

By (3.18), we have

$$(N+n)^{1/2}\widetilde{U}_{nN}(\beta) = \{\lambda/(\lambda+1)\}^{1/2} n^{-1} N^{-1/2} \textstyle\sum_{i=1}^{N} \sum_{j=N+1}^{N+n} \widetilde{\psi}_n(U_i, \widetilde{V}_j; \beta) + o_p(1),$$

where $U_i = (Y_i, W_i)$, $\widetilde{V}_j = (Y_j, X_j, W_j)$, and $\widetilde{\psi}_n(\beta; U_i, \widetilde{V}_j) = \psi_n(\beta; U_i, V_j) + \lambda^{-1} U(\beta; X_j, Y_j)$. Similarly, we can show that $E\{\widetilde{\psi}_n(\beta; U_i, \widetilde{V}_j) \mid U_i\} = e(\beta; Y_i, W_i)$, the estimating equation (3.6) is asymptotically unbiased, and $E\{\widetilde{\psi}_n(\beta; U_i, \widetilde{V}_j) \mid \widetilde{V}_j\} = M(\beta; X_j, W_j) + \lambda^{-1} U(\beta; X_j, Y_j) + O_p(h^k)$. Similar to the proof of Theorem 1, we have $(N+n)^{1/2}\widetilde{U}_{nN}(\beta) \xrightarrow{D} N(0, \Sigma_2)$, and $\widetilde{J}_{nN}(\beta^*) \xrightarrow{p} \Gamma_2$, and then Theorem 2 follows from Slutsky's theorem.

CHAPTER 4. NESTED HIERARCHICAL FUNCTIONAL DATA MODELING FOR ROOT GRAVITROPISM DATA AND TESTING FOR MOON PHASE EFFECT

A paper to be submitted to *Journal of the American Statistical Association*

Yuhang Xu[1], Yehua Li[2] Dan Nettleton[3]

Abstract

In a Root Image Study in plant science, the root bending process of seeds from various genotypes are recorded using digital cameras, and the bending rates are modeled as functional data. The data are collected from seeds with a large variety of genotypes and have a three-level nested hierarchical structure – multiple seeds from the same genotype are recorded using the same protocol under different camera setups. The seeds are planted on different lunar days and an important scientific question is whether the moon phase has any effect on root bending. We allow the mean function of the root bending rate to depend on the lunar day and model the variation between genotypes, camera files and seeds by hierarchical functional random effects. We estimate the covariance functions of the functional random effects by a fast penalized tensor product spline approach, perform multi-level functional principal component analysis (FPCA) using the best linear unbiased predictor of the principal component scores, and improve the efficiency of

[1]Primary researcher and author, Graduate student, Department of Statistics, Iowa State University.
[2]Author for correspondence, Associate Professor, Department of Statistics, Iowa State University.
[3]Distinguished Professor, Department of Statistics, Iowa State University.

mean estimation by iterative decorrelation. We choose the number of principal components using a conditional Akaike Information Criterion and test the lunar day effect using generalized likelihood ratio test statistics based on the marginal and conditional likelihoods. We also propose a permutation procedure to evaluate the null distribution of the test statistics. Our simulation studies show that our model selection criterion selects the correct number of principal components with remarkable high frequency, and the likelihood-based tests based on FPCA have higher power than the test based on working independence. We have also discovered a significant moon phase effect in our real data.

4.1 Introduction

Gravitropism is the growth movement of a plant in response to gravity. Charles Darwin was one of the first scientists to document that plant roots show positive gravitropism, i.e. growing in the same direction as gravity. Root gravitropism is a very active research area in botany and agriculture because of its importance for plant growth and development (Noh et al., 2003). In a recent Root Image Study (RIS) conducted by Edgar Spalding's lab at the University of Wisconsin-Madison, researchers studied root gravitropism of maize seeds with various genotypes. There were 1762 seeds observed in the RIS, drawn from 235 genotypes with up to 10 replicates for each genotype. Within each genotype, seeds were planted in up to two dishes. Each dish contained up to 5 seeds and was monitored by one digital camera. Figure 4.1 (a) shows an image of a group of seeds of the same genotype, planted in the same dish. We refer to a dish of seeds as a 'file' in this paper because images of these seeds were recorded in one camera file. There were a total of 457 files in the RIS data. The lab had 7 cameras and could monitor multiple dishes the same day, and the whole experiment was completed in about 4 months. The data have a natural three-level nested hierarchical structure, with files (level two) nested in genotypes (level one) and seeds (level three) nested in files.

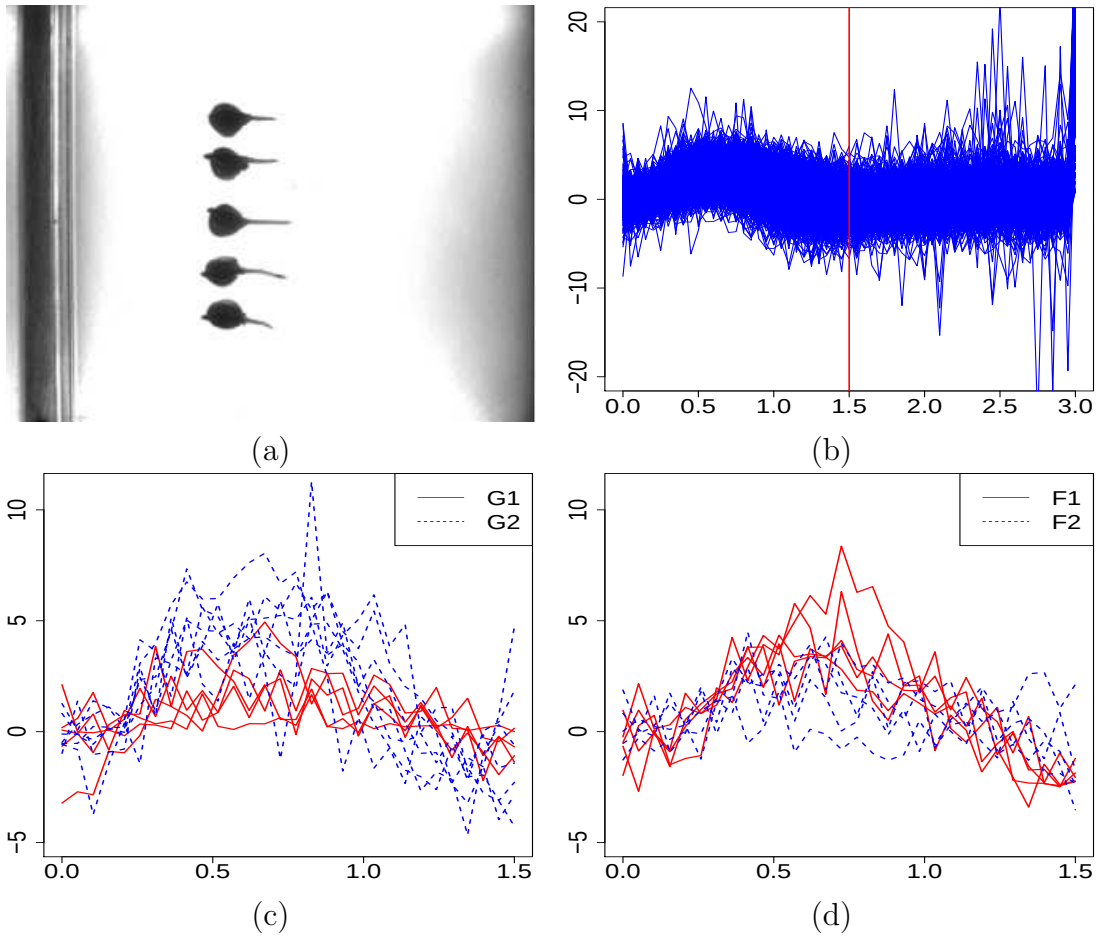


Figure 4.1 The root gravitropism data. Panel (a) shows the image of a group of seeds under an experimental setup; panel (b) shows the bending rate process for all seeds during the 3-hour experiment time; panel (c) shows the process for seeds from two randomly selected genotypes in the first 1.5 hours; panel (d) shows the process for seeds from the two files within the same genotype in the first 1.5 hours.

When the seeds were planted, their root tips were close to being horizontal, as shown in Figure 4.1 (a). As the seeds sprouted, their root tips turned downward due to root gravitropism. The root tip angle of each seed, with respect to the horizontal line, was recorded by a camera every 3 minutes for a total duration of 3 hours. To get rid of the nuisance from the initial root tip angle, we take differences between the root tip angles at adjacent time points and the resulting response variable is the bending rate of a root tip. Figure 4.1 (b) shows the bending rates for all maize seeds. As shown in the plot, root tip bending mostly occurs during the first 1.5 hours, therefore we keep the most informative part of the data and only model the bending rate process within the first 1.5 hours. Figure 4.1 (c) illustrates the variation between two randomly selected genotypes and panel (d) illustrates the variation between two files within the same genotype. The bending rate is a time-dependent process and hence naturally modeled as functional data (Ramsay and Silverman, 2005), but these functions have a nested correlation structure inherited from the experimental design. We model the genotype, file and seed effects as nested functional random effects, and each can be represented by a Karhunen-Loève expansion. This approach is often referred to as hierarchical or multi-level functional principal component analysis (FPCA).

Over the 4 months duration of the study, the seeds were planted on different lunar days corresponding to different moon phases. It is well-known that in many cultures, agricultural activities such as planting and harvesting are scheduled according to moon phases. Scientists want to know if there is any wisdom in these traditions that can be backed up by science. We know moon phase is connected with the distance between the moon and the earth, which affects the gravity on earth. It is of scientific interest to model and test moon phase effects on root tip bending of maize seeds. We model the bending rate trajectories of the seeds as hierarchical functional data, allowing the mean function of the bending rate to depend on both the lunar day that they were planted on and the time since the seed was planted.

In functional data analysis (FDA), data are usually curves or images. FPCA has become one of the most important modeling and dimension reduction tools in FDA. Some classic work on FPCA methodology includes Yao and Lee (2006) and James, Hastie, and Sugar (2000), and the theoretical properties of FPCA are investigated by Hall, Müller, and Wang (2006) and Li and Hsing (2010). However, these papers only study samples of independent curves. As technology advances, hierarchical functional data become increasingly available. Di et al. (2009) studies two-level hierarchical functional data from a sleep heart health study, where each subject yields multiple electroencephalographic (EEG) curves from multiple hospital visits. Li, et al. (2015) analyzes three-level functional data from an exercise intervention trial where real time measurements on the activity level (measured by metabolic units or METs) have a subject-week-day three-level hierarchical structure. Other related papers include Zhou, Huang, and Carroll (2008); Zhou et al. (2010) and Serban and Jiang (2012).

Compared with existing methodology on analyzing hierarchical functional data, our main contributions are the following. First, we estimate the mean function by anisotropic bivariate penalized splines and adopt a fast hierarchical FPCA algorithm, which directly estimates the covariance functions of the functional random effects using a method-of-moment approach based on penalized tensor product B-spline smoothing (Ruppert, Wand, Carroll, 2003). Our algorithm can handle large functional data sets and does not involve computationally intensive EM iterations as in Li, et al. (2015). Compared with Di et al. (2009), we provide more detailed smoothing strategies to eliminate measurement errors, and estimate the principal component scores using the best linear unbiased predictor (BLUP) method rather than time consuming Markov Chain Monte Carlo (MCMC). To improve the estimation efficiency of the mean function, we adopt an iterative decorrelation procedure similar to that of Yao and Lee (2006) for uni-level functional data.

Second, we propose a new method to choose the number of principal components based on a conditional Aikaike information criterion (AIC). Selecting the number of

principal components is one of the most important model selection issues in FPCA. The current literature on hierarchical FPCA, including Di et al. (2009) and Li, et al. (2015), selects the number of components subjectively using an ad hoc "percentage of variation explained" (PVE) method. In contrast, our proposed method, which extends the recent work of Li, Wang, and Carroll (2013) for independent functional data to the hierarchical setting, is completely data-driven and vastly outperforms the existing ad hoc methods in our simulations studies.

Third, and most importantly, we propose new test procedures on the mean function based on generalized likelihood ratio (GLR) test statistics (Fan, Zhang, and Zhang, 2001). There is relatively little work on nonparametric inference for hierarchical functional data, with the exception of Li, et al. (2015) which considers a Wald test on the mean parameters. In our model, in order to test the moon phase effect, we compare two models: in the full model, the mean bending rate is a bivariate function that depends on both the recording time since the seed being planted and the lunar day; in the reduced model, the mean function is univariate and does not depend on the lunar day. To the best of our knowledge, nonparametric tests on bivariate alternative versus univariate null have not been investigated in the literature. We propose three versions of GLR tests based on the marginal likelihood, conditional likelihood and working independence, respectively. We propose a simple permutation strategy to estimate the null distribution of these test statistics, and investigate the empirical size and power of the proposed test procedures.

The rest of paper is organized as follows. We describe the hierarchical functional data model in Section 4.2 and the estimation procedure in Section 4.3. We address the model selection and inference issues in Section 4.4, illustrate the proposed methods by simulation studies in Section 4.5, and analyze our motivating data set in Section 4.6. Some concluding remarks are provided in Section 4.7. The online supplementary material contains additional simulation results and graphs.

4.2 Hierarchical functional data modeling

Let $Y_i(t)$ be the bending rate of the ith seed at time $t \in \mathcal{T}$, where $\mathcal{T} = [0, 1.5]$ is the observation time period and $i = 1, 2, \ldots, n$. Denote s_i as the lunar day on which the ith seed was planted, where $s_i \in \mathbb{S}$ and $\mathbb{S} = [1, 30]$. For each seed, we also observe a covariate vector $\mathbf{X}_i$. In the RIS data, $\mathbf{X}_i$ contains indicators of different camera setups, the effects of which are modeled as fixed effects that do not drift over time. Since the data are collected from a large number of genotypes, and seeds are measured in groups (i.e. files), the effects of these factors are random and evolve over time. There are a total of G genotypes documented in F files. We use g, f and i for indices of genotype, file and seed respectively. With a slight abuse of notation, we also use $g(\cdot)$ and $f(\cdot)$ as index functions, e.g. $g(i)$ and $f(i)$ are the genotype and file number of the ith seed respectively. For the gth genotype, define $n_g = \#\{i : g(i) = g\}$ as the number of seeds with genotype g and $F_g = \#\{f : g(f) = g\}$ as the number of files with genotype g.

We model the RIS data by the following hierarchical functional data model

$$Y_i(t) = \mu(s_i, t) + \mathbf{X}_i'\boldsymbol{\alpha} + Z_{1,g(i)}(t) + Z_{2,f(i)}(t) + Z_{3,i}(t) + \epsilon_i(t), \tag{4.1}$$

where $\mu(s_i, t)$ is the mean function of the bending rate under a baseline camera setup, $\boldsymbol{\alpha}$ represents fixed effects of cameras, $Z_{1,g}(t)$, $Z_{2,f}(t)$, and $Z_{3,i}(t)$ are random processes representing the functional random effects of genotype g, file f, and seed i, respectively, and $\epsilon_i(t)$ is a white noise measurement error with variance σ^2. We assume that $Z_l(t)$, $l = 1, 2, 3$, are zero-mean random processes in time with covariance functions

$$\mathcal{K}_l(t_1, t_2) = \mathrm{Cov}\{Z_l(t_1), Z_l(t_2)\}. \tag{4.2}$$

Furthermore, we assume that $Z_1(t)$, $Z_2(t)$, $Z_3(t)$ and $\epsilon(t)$ are mutually independent.

The covariance functions in (4.2) are positive semi-definite bivariate functions with the following spectral decomposition

$$\mathcal{K}_l(t_1, t_2) = \sum_{k=1}^{\infty} \omega_{l,k}\psi_{l,k}(t_1)\psi_{l,k}(t_2), \quad l = 1, 2, 3, \tag{4.3}$$

where $\omega_{l,k}$ are the eigenvalues of $\mathcal{K}_l$ in a descending order and $\psi_{l,k}$ are the corresponding eigenfunctions. The eigenfunctions are orthonormal in the sense that $\int_{\mathcal{T}} \psi_{l,k}(t)\psi_{l,k'}(t)\mathrm{d}t$ equals to 1 if $k = k'$ and equals to 0 otherwise. By the Karhunen-Loève expansion,

$$Z_l(t) = \sum_{k=1}^{\infty} \xi_{l,k}\psi_{l,k}(t), \tag{4.4}$$

where $\xi_{l,k}$ are zero-mean, uncorrelated random variables, known as the principal component scores of Z_l such that $\mathrm{Var}(\xi_{l,k}) = \omega_{l,k}$. In practice, the Karhunen-Loève expansions in (4.4) need to be truncated at finite orders. Suppose that the process $Z_l(t)$ can be characterized by p_l principal components, $l = 1, 2, 3$. These numbers determine the model complexity of the longitudinal correlation structure, so selection of these numbers plays a key role in FPCA. We will propose a data-driven model selection method in Section 4.4.1 to choose these numbers.

4.3 Estimation procedure

4.3.1 Estimating the mean and covariance functions

We first estimate the mean function $\mu(s,t)$ by penalized tensor product spline regression. In the real data, we have $m_T = 31$ observation points in $\mathcal{T}$ and $m_S = 30$ time points in $\mathbb{S}$. Notice that the moon phase effect is periodic by nature with a 30-day cycle but the effect in t does not have such a constraint. We therefore use different basis functions in the two domains. Define B-spline basis functions $\mathbf{B}_T(t) = (B_{T1}, \ldots, B_{TK_T})'(t)$ on $\mathcal{T}$ with equally-spaced interior knots, and let $\mathbf{B}_L(s) = (B_{L1}, \ldots, B_{LK_L})'(s)$ be Fourier basis functions with a 30-day period on $\mathbb{S}$. Define the tensor product basis on $\mathbb{S} \times \mathcal{T}$ as $\mathbf{B}_\mu(s,t) = \mathbf{B}_L(s) \otimes \mathbf{B}_T(t)$ where $\otimes$ is the Kronecker product. Then we can approximate $\mu(s,t)$ by $\mathbf{B}'_\mu(s,t)\boldsymbol{\beta}_\mu$ and estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_\mu$ by minimizing the following penalized sum of squares

$$\sum_{i=1}^{n}\sum_{j=1}^{m_T}\{Y_i(t_j) - \mathbf{B}'_\mu(s_i,t_j)\boldsymbol{\beta}_\mu - \mathbf{X}'_i\boldsymbol{\alpha}\}^2 + \mathcal{P}(\boldsymbol{\beta}_\mu; \lambda_\mu, \varrho), \tag{4.5}$$

where $\mathcal{P}(\boldsymbol{\beta}_\mu; \lambda_\mu, \varrho)$ is a penalty on $\boldsymbol{\beta}_\mu$ with tuning parameters λ_μ and ϱ. To increase model flexibility, we allow μ to have different degrees of smoothness in s and t by introducing two tuning parameters. To mimic the anisotropic thin-plate spline (Wood, 2000), we put penalty on

$$\lambda_\mu \int_{\mathbb{S}} \int_{\mathcal{T}} [\{\mu^{(2,0)}(s,t)\}^2 + 2\varrho\{\mu^{(1,1)}(s,t)\}^2 + \varrho^3\{\mu^{(0,2)}(s,t)\}^2]\, \mathrm{d}t\mathrm{d}s$$

where $\mu^{(k_1,k_2)}$ is the (k_1,k_2)th partial derivative of μ. Using the basis function representation, the thin-plate penalty can be written as

$$\mathcal{P}(\boldsymbol{\beta}_\mu; \lambda_\mu, \varrho) = \lambda_\mu \boldsymbol{\beta}_\mu' \int_{\mathbb{S}} \int_{\mathcal{T}} [\{\mathbf{B}_\mu^{(2,0)}(s,t)\}^{\otimes 2} + 2\varrho\{\mathbf{B}_\mu^{(1,1)}(s,t)\}^{\otimes 2} + \varrho^3\{\mathbf{B}_\mu^{(0,2)}(s,t)\}^{\otimes 2}]\, \mathrm{d}t\mathrm{d}s\boldsymbol{\beta}_\mu,$$

where $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}'$ for any matrix $\mathbf{A}$. Following Ruppert, Wand, Carroll (2003), we set both K_T and K_L to be relatively large and let the smoothness of the estimated function controlled by the tuning parameters (λ_μ, ϱ), which can be selected by data-driven methods such as the generalized cross-validation (GCV) (Wahba, 1990). Denote the estimators for camera effects and the mean function as $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\mu}(s,t)$.

Under model (4.1), we can easily see

$$\begin{aligned}
&\mathcal{G}_1(t_1,t_2) \equiv \mathrm{Cov}\{Y_i(t_1), Y_i(t_2)\} = \mathcal{K}_1(t_1,t_2) + \mathcal{K}_2(t_1,t_2) + \mathcal{K}_3(t_1,t_2), \text{ if } t_1 \neq t_2;\\
&\mathcal{G}_2(t_1,t_2) \equiv \mathrm{Cov}\{Y_{i_1}(t_1), Y_{i_2}(t_2)\}\\
&\qquad\qquad = \mathcal{K}_1(t_1,t_2) + \mathcal{K}_2(t_1,t_2), \text{ if } i_1 \neq i_2,\ g(i_1) = g(i_2),\ f(i_1) = f(i_2);\\
&\mathcal{G}_3(t_1,t_2) \equiv \mathrm{Cov}\{Y_{i_1}(t_1), Y_{i_2}(t_2)\} = \mathcal{K}_1(t_1,t_2), \text{ if } i_1 \neq i_2,\ g(i_1) = g(i_2),\ f(i_1) \neq f(i_2);\\
&\sigma_Y^2(t) \equiv \mathrm{Var}\{Y_i(t)\} = \mathcal{K}_1(t,t) + \mathcal{K}_2(t,t) + \mathcal{K}_3(t,t) + \sigma_\epsilon^2.
\end{aligned}$$

We will first estimate $\mathcal{G}_l$, $l = 1,2,3$, and then use the above relationships to estimate $\mathcal{K}_l$.

To estimate the covariance functions, we first define the residuals $\mathcal{E}_{ij} = Y_i(t_j) - \widehat{\mu}(s_i,t_j) - \mathbf{X}_i'\widehat{\boldsymbol{\alpha}}$. The equations above suggest that we can estimated $\mathcal{G}_1(t_{j_1}, t_{j_2})$ by $\widehat{\mathcal{G}}_1(t_{j_1}, t_{j_2}) = \frac{1}{n}\sum_{i=1}^n \mathcal{E}_{ij_1}\mathcal{E}_{ij_2}$ for $j_1 \neq j_2$. However, these estimates are only defined on discrete time points with 3-minute gaps between consecutive time points. To estimate

$\mathcal{G}_1$ as a function, we apply penalized tensor-product spline smoothing to these empirical covariance estimators. Define the second tensor-product spline basis $\mathbf{B}_{\mathcal{G}}(t_1,t_2) = \mathbf{B}_T(t_1)\otimes\mathbf{B}_T(t_2)$, then we approximate $\mathcal{G}_1(t_1,t_2)$ by $\widehat{\mathcal{G}}_1(t_1,t_2) = \mathbf{B}'_{\mathcal{G}}(t_1,t_2)\widehat{\boldsymbol{\beta}}_{\mathcal{G}1}$, where $\widehat{\boldsymbol{\beta}}_{\mathcal{G}1}$ minimizes

$$\sum_{i=1}^{n}\sum_{j_1=1}^{m_T}\sum_{j_2\neq j_1}\{\mathcal{E}_{ij_1}\mathcal{E}_{ij_2} - \mathbf{B}'_{\mathcal{G}}(t_{j_1},t_{j_2})\boldsymbol{\beta}_{\mathcal{G}1}\}^2 + \lambda_{\mathcal{G}1}\boldsymbol{\beta}'_{\mathcal{G}1}\boldsymbol{\Omega}_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}1},$$

and $\lambda_{\mathcal{G}1}$ and $\boldsymbol{\Omega}_{\mathcal{G}}$ are the tuning parameter and penalty matrix respectively.

Similarly, we estimate $\mathcal{G}_2$ and $\mathcal{G}_3$ by $\widehat{\mathcal{G}}_2 = \mathbf{B}'_{\mathcal{G}}\widehat{\boldsymbol{\beta}}_{\mathcal{G}2}$ and $\widehat{\mathcal{G}}_3 = \mathbf{B}'_{\mathcal{G}}\widehat{\boldsymbol{\beta}}_{\mathcal{G}3}$ where $\widehat{\boldsymbol{\beta}}_{\mathcal{G}2}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{G}3}$ minimize the following penalized sum of squares

$$\begin{aligned}
&\sum_{i_1=1}^{n}\sum_{i_2\neq i_1}\sum_{j_1=1}^{m_T}\sum_{j_2=1}^{m_T}\{\mathcal{E}_{i_1j_1}\mathcal{E}_{i_2j_2} - \mathbf{B}'_{\mathcal{G}}(t_{j_1},t_{j_2})\boldsymbol{\beta}_{\mathcal{G}2}\}^2 I\{g(i_1)=g(i_2), f(i_1)=f(i_2)\}\\
&+\lambda_{\mathcal{G}2}\boldsymbol{\beta}'_{\mathcal{G}2}\boldsymbol{\Omega}_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}2},\\
&\sum_{i_1=1}^{n}\sum_{i_2\neq i_1}\sum_{j_1=1}^{m_T}\sum_{j_2=1}^{m_T}\{\mathcal{E}_{i_1j_1}\mathcal{E}_{i_2j_2} - \mathbf{B}'_{\mathcal{G}}(t_{j_1},t_{j_2})\boldsymbol{\beta}_{\mathcal{G}3}\}^2 I\{g(i_1)=g(i_2), f(i_1)\neq f(i_2)\}\\
&+\lambda_{\mathcal{G}3}\boldsymbol{\beta}'_{\mathcal{G}3}\boldsymbol{\Omega}_{\mathcal{G}}\boldsymbol{\beta}_{\mathcal{G}3}.
\end{aligned}$$

When $\mathcal{G}$ is spanned by a tensor-product spline basis $\mathbf{B}_{\mathcal{G}}$, the penalty matrix corresponding to a thin-plate spline penalty is

$$\boldsymbol{\Omega}_{\mathcal{G}} = \int_{\mathcal{T}}\int_{\mathcal{T}}\{\mathbf{B}^{(2,0)}_{\mathcal{G}}(t_1,t_2)\}^{\otimes 2} + 2\{\mathbf{B}^{(1,1)}_{\mathcal{G}}(t_1,t_2)\}^{\otimes 2} + \{\mathbf{B}^{(0,2)}_{\mathcal{G}}(t_1,t_2)\}^{\otimes 2}\mathrm{d}t_1\mathrm{d}t_2.$$

We can also estimate $\sigma^2_Y(t)$ by $\widehat{\sigma}^2_Y(t) = \mathbf{B}'_T(t)\widehat{\boldsymbol{\beta}}_{\sigma}$, where $\widehat{\boldsymbol{\beta}}_{\sigma}$ minimizes

$$\sum_{i=1}^{n}\sum_{j=1}^{m_T}\{\mathcal{E}^2_{ij} - \mathbf{B}_T(t_j)\boldsymbol{\beta}_{\sigma}\}^2 + \lambda_{\sigma}\boldsymbol{\beta}_{\sigma}\boldsymbol{\Omega}_T\boldsymbol{\beta}_{\sigma},$$

and λ_{σ} and $\boldsymbol{\Omega}_T = \int\{\mathbf{B}^{(2)}_T(t)\}^{\otimes 2}\mathrm{d}t$ are the tuning parameter and penalty matrix respectively. All tuning parameters defined above can be chosen by GCV.

Next, we estimate the covariance functions $\mathcal{K}_l, l = 1,2,3$ and the error variance σ^2 by

$$\begin{aligned}
&\widehat{\mathcal{K}}_1(t_1,t_2) = \widehat{\mathcal{G}}_3(t_1,t_2), \quad \widehat{\mathcal{K}}_2(t_1,t_2) = \widehat{\mathcal{G}}_2(t_1,t_2) - \widehat{\mathcal{G}}_3(t_1,t_2),\\
&\widehat{\mathcal{K}}_3(t_1,t_2) = \widehat{\mathcal{G}}_1(t_1,t_2) - \widehat{\mathcal{G}}_2(t_1,t_2),\\
&\widehat{\sigma}^2_I = |\mathcal{T}|^{-1}\int\{\widehat{\sigma}^2_Y(t) - \widehat{\mathcal{K}}_1(t,t) - \widehat{\mathcal{K}}_2(t,t) - \widehat{\mathcal{K}}_3(t,t)\}\mathrm{d}t. \qquad (4.6)
\end{aligned}$$

The eigenvalues and eigenfunctions in (4.3) can be estimated by an eigenvalue decomposition of $\widehat{\mathcal{K}}_l$ using the approach of Ramsay and Silverman (2005). Since all functions above are approximated by finite-dimensional splines, the eigenvalue decomposition problem reduces to a multivariate problem.

4.3.2 Estimating the principal component scores

For predetermined numbers of principal components (p_1, p_2, p_3) for the three levels of functional random effects, we estimate the principal component scores by best linear unbiased prediction (BLUP). This is an extension of the "PACE" method of Yao, Müller, and Wang (2005) to hierarchical functional data.

Define the following notations: $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,m_T})'$, $\boldsymbol{\mu}_i = \{\mu(s_i, t_1), \ldots, \mu(s_i, t_{m_T})\}'$, $\boldsymbol{\psi}_{l,k} = \{\psi_{l,k}(t_1), \ldots, \psi_{l,k}(t_{m_T})\}'$, $\boldsymbol{\Psi}_l = (\boldsymbol{\psi}_{l,1}, \boldsymbol{\psi}_{l,2}, \ldots, \boldsymbol{\psi}_{l,p_l})$, $l = 1, 2, 3$. Define the vectors of FPCA scores $\boldsymbol{\xi}_{1,g} = \{\xi_{1,g,1}, \ldots, \xi_{1,g,p_1}\}'$ for $g = 1 \ldots, G$, $\boldsymbol{\xi}_{2,f} = \{\xi_{2,f,1}, \ldots, \xi_{2,f,p_2}\}'$ for $f = 1, \ldots, F$, and $\boldsymbol{\xi}_{3,i} = \{\xi_{3,i,1}, \ldots, \xi_{3,i,p_3}\}'$ for $i = 1, \ldots, n$. For any positive integer d, denote $\mathbf{1}_d$ as a d-dimensional vector of ones and $\mathbf{I}_d$ as a d by d identity matrix. For any index set $I = \{k_1, \ldots, k_d\} \subset \{1, \ldots, n\}$ and any sequence of vectors or matrices $\mathbf{A}_k$ of the same size, denote $(\mathbf{A}_k)_{k \in I}$ as $(\mathbf{A}'_{k_1}, \ldots, \mathbf{A}'_{k_d})'$. For $g = 1, \ldots, G$, define $\boldsymbol{Y}_g = (\boldsymbol{Y}_i)_{g(i)=g}$, $\boldsymbol{\mu}_g = (\boldsymbol{\mu}_i)_{g(i)=g}$, $\mathbf{X}_g = (\mathbf{1}_{m_T} \otimes \mathbf{X}'_i)_{g(i)=g}$, $\boldsymbol{\xi}_g = \{\boldsymbol{\xi}'_{1,g}, (\boldsymbol{\xi}_{2,f})'_{g(f)=g}, (\boldsymbol{\xi}_{3,i})'_{g(i)=g}\}'$, $\boldsymbol{\Lambda}_g = \mathrm{diag}\{\mathrm{Var}(\boldsymbol{\xi}_g)\}$, $\boldsymbol{\Phi}_g = (\mathbf{1}_{n_g} \otimes \boldsymbol{\Psi}_1, \mathbf{D}_g \otimes \boldsymbol{\Psi}_2, \mathbf{I}_{n_g} \otimes \boldsymbol{\Psi}_3)$, where $\mathbf{D}_g$ is a n_g by F_g matrix with its (k_1, k_2) element equals to 1 if the k_1th seed in genotype g is recorded in the k_2th file and 0 otherwise. It is easy to see that $\mathrm{Cov}(\boldsymbol{Y}_g) = \boldsymbol{\Sigma}_g$ where $\boldsymbol{\Sigma}_g = \boldsymbol{\Omega}_g + \sigma^2 \mathbf{I}_{m_T n_g}$ where $\boldsymbol{\Omega}_g = \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g \boldsymbol{\Phi}'_g$. Under the assumption that $\boldsymbol{\xi}_g$ and $\epsilon(t)$ are jointly Gaussian, $\mathrm{E}(\boldsymbol{\xi}_g | \boldsymbol{Y}_g, \mathbf{X}_g) = \boldsymbol{\Lambda}_g \boldsymbol{\Phi}'_g \boldsymbol{\Sigma}_g^{-1} (\boldsymbol{Y}_g - \boldsymbol{\mu}_g - \mathbf{X}_g \boldsymbol{\alpha})$. The estimator of $\boldsymbol{\xi}_g$ is its empirical BLUP

$$\widehat{\boldsymbol{\xi}}_g = \widehat{\boldsymbol{\Lambda}}_g \widehat{\boldsymbol{\Phi}}'_g \widehat{\boldsymbol{\Sigma}}_g^{-1} (\boldsymbol{Y}_g - \widehat{\boldsymbol{\mu}}_g - \mathbf{X}_g \widehat{\boldsymbol{\alpha}}),$$

where $\widehat{\boldsymbol{\mu}}_g$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\Lambda}}_g$, and $\widehat{\boldsymbol{\Phi}}_g$ are the estimates using the proposed FPCA method described in Section 4.3.1, $\widehat{\boldsymbol{\Sigma}}_g = \widehat{\boldsymbol{\Phi}}_g \widehat{\boldsymbol{\Lambda}}_g \widehat{\boldsymbol{\Phi}}'_g + \widehat{\sigma}_I^2 \mathbf{I}_{m_T n_g}$, and $\widehat{\sigma}_I^2$ is a pilot estimator of σ^2 obtained by integration defined in (4.6).

4.3.3 Iterative procedure to refine mean estimation

The algorithm we propose in Section 4.3.1 to estimate the mean and covariance functions is an extension of the method for two-level hierarchical functional data in Di et al. (2009) to the three-level setting but with more emphasis on smoothing. The benefit of this approach is that it does not involve computationally intensive EM algorithms as in Li, et al. (2015) and can handle large functional data sets. However, the estimator in (4.5) is a working independence estimator (Lin and Carroll, 2001) which ignores correlation in the data, so it is not efficient. To improve the estimation efficiency and increase the power for statistical tests, we refine the mean estimator using an iterative decorrelation procedure similar to that of Yao and Lee (2006) for uni-level FPCA.

The iterative procedure is as follows.

Step 1: Use the procedures in Section 4.3.1 to obtain $\widehat{\boldsymbol{\mu}}_i$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\mathcal{K}}_l$, $l = 1, 2, 3$, and perform spectral decomposition to these covariance functions.

Step 2: Use AIC defined in Section 4.4.1 to choose the numbers of principal components $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3)$ for the three levels and obtain estimates of the eigenvalues $\widehat{\boldsymbol{\Lambda}}_g$, eigenfunctions $\widehat{\boldsymbol{\Psi}}_g$ and the principal component scores $\widehat{\boldsymbol{\xi}}_g$, $g = 1, \ldots, G$.

Step 3: Re-estimate $\mu(s,t)$ and $\boldsymbol{\alpha}$ using (4.5) while replace $Y_i(t_j)$ by

$$Y_i^*(t_j) = Y_i(t_j) - \sum_{k=1}^{\widehat{p}_1} \widehat{\xi}_{1,g(i),k}\widehat{\psi}_{1,k}(t_j) - \sum_{k=1}^{\widehat{p}_2} \widehat{\xi}_{2,f(i),k}\widehat{\psi}_{2,k}(t_j) - \sum_{k=1}^{\widehat{p}_3} \widehat{\xi}_{3,i,k}\widehat{\psi}_{3,k}(t_j). \quad (4.7)$$

Step 4: Update the covariance estimates using the updated mean function and parameters; update the estimates of the eigenvalues, eigenfunctions, and principal component scores; and if necessary, adjust the numbers of principal components using AIC.

Step 5: Repeat *Step 3* and *Step 4* until relative changes in $\widehat{\mu}$ and $\widehat{\boldsymbol{\alpha}}$ between adjacent iterations are smaller than a predetermined tolerance. The final numbers of principal components are the ones picked by AIC in the final step.

The decorrelation procedure in *Step 3* is an extension of the algorithm in Yao and Lee (2006) to hierarchical functional data, where we get rid of the correlation in the response by subtracting the predicted functional random effects. The estimator in (4.5) is efficient for uncorrelated responses.

In our experience, the numbers of principal components are usually chosen perfectly by AIC in *Step 2*. To save computation time, the subsequent updates of $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3)$ can be skipped in *Step 4*.

4.4 Model selection and statistical inference

4.4.1 Selecting the number of principal components

All existing papers on hierarchical functional data analysis select the number of principal components using the ad hoc PVE method, where the estimated eigenvalue sequence of each functional random effect is truncated at a subjectively chosen percentage of variation explained. More recently, Li, Wang, and Carroll (2013) proposed a model selection method using conditional AIC to select the number of principal components for uni-level functional data, and we now extend their method to hierarchical functional data.

Assuming the measurement errors are Gaussian, the conditional log-likelihood of the observed data $\{\boldsymbol{Y}_i\}_{i=1}^n$ given the principal component scores is

$$l_{n,\mathcal{C}} = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\|\boldsymbol{Y}_i - \boldsymbol{\mu}_i - \mathbf{X}_i'\boldsymbol{\alpha} - \boldsymbol{\Psi}_1\boldsymbol{\xi}_{1,g(i)} - \boldsymbol{\Psi}_2\boldsymbol{\xi}_{2,f(i)} - \boldsymbol{\Psi}_3\boldsymbol{\xi}_{3,i}\|^2, \quad (4.8)$$

where $\|\cdot\|$ denotes the L^2 norm and $N = nm_T$ is the total number of measurements.

For a given model specified by the numbers of components (p_1, p_2, p_3), we first predict the functional random effects using the procedure in Section 4.3.2 and evaluate the conditional likelihood of the observed data by replacing various fixed and random effects by their estimators or predictors. The AIC is the value of negative conditional log-likelihood plus a penalty on the complexity of the hierarchical functional data model.

Motivated by Li, Wang, and Carroll (2013), since the functional random processes for the three levels are mutually independent, the penalty is on the number of estimated random effects which yields the following AIC

$$\text{AIC}(p_1, p_2, p_3) \quad = \quad -2\widehat{l}_{n,\mathcal{C}} + 2(Gp_1 + Fp_2 + np_3), \tag{4.9}$$

where $\widehat{l}_{n,\mathcal{C}}$ is the estimated conditional likelihood as described above. The numbers of components are selected by minimizing (4.9) using a grid search method.

An intuitive explanation for the penalty in (4.9) is given as follows. The fixed effects including $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\Psi}_1$, $\boldsymbol{\Psi}_2$ and $\boldsymbol{\Psi}_3$ are estimated with the highest accuracy by pooling all data together and hence can be deemed as known. Then the likelihood in (4.8) can be considered as a regression on $\boldsymbol{Y}_i - \boldsymbol{\mu}_i - \mathbf{X}_i'\boldsymbol{\alpha}$ against $\boldsymbol{\Psi}_1$, $\boldsymbol{\Psi}_2$ and $\boldsymbol{\Psi}_3$, where $\boldsymbol{\xi}_{1,g}$, $\boldsymbol{\xi}_{2,f}$ and $\boldsymbol{\xi}_{3,i}$ are the genotype-, file- and seed-specific regression coefficients. The total number of regression coefficients is $Gp_1 + Fp_2 + np_3$. This calculation is logical because we have dense observations on each curve so that there are enough data to fit a regression in each seed.

4.4.2 Test on moon phase effect

To test the moon phase effect, we consider a reduced model of (4.1), where the mean of the response $Y_i(t)$ does not depend on the lunar day s, i.e. $\mu(s,t) \equiv \mu_R(t)$. We will test the hypothesis

$$H_0 : \mu(s,t) = \mu_R(t) \quad \text{vs.} \quad H_1 : \mu(s,t) \neq \mu_R(t), \tag{4.10}$$

by a generalized likelihood ratio (GLR) test (Fan, Zhang, and Zhang, 2001). The classic GLR test was proposed for testing nonparametric hypotheses for independent data. Some recent reviews on this test include Fan and Jiang (2007), and González-Manteiga and Crujeiras (2013). It is also recently extended to uni-level sparse functional data by Tang, Li, and Guan (2016), who build a GLR test based on working independent estimators.

In our setting, we introduce three versions of GLR tests based on marginal likelihood, conditional likelihood and working independence (WI), respectively.

Under the full model and Gaussian assumptions, the estimated marginal likelihood is

$$\widehat{l}_{n,\mathcal{M}}(H_1) \propto -\frac{1}{2}\sum_{g=1}^{G}(\boldsymbol{Y}_g - \widehat{\boldsymbol{\mu}}_g - \mathbf{X}_g\widehat{\boldsymbol{\alpha}})'\widehat{\boldsymbol{\Sigma}}_g^{-1}(\boldsymbol{Y}_g - \widehat{\boldsymbol{\mu}}_g - \mathbf{X}_g\widehat{\boldsymbol{\alpha}}), \tag{4.11}$$

where $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\alpha}}$ are the refined estimators in Section 4.3.3, $\widehat{\boldsymbol{\Sigma}}_g$ is the covariance matrix for the response variables within genotype g reconstructed from FPCA as in Section 4.3.2 using the selected numbers of components.

Since the reduced model is nested in the full model, the FPCA estimators (including the eigenvalues, eigenfunctions and FPC scores) under the full model are still legitimate when H_0 is true. We fit $\mu_R(t)$ using univariate penalized splines to the decorrelated response in (4.7), where the principal component scores, eigenvalues, and eigenfunctions are obtained from the full model. The likelihood under the reduced model, denoted as $\widehat{l}_{n,\mathcal{M}}(H_0)$, is similarly defined as in (4.11), where $\widehat{\boldsymbol{\Sigma}}_g$ are obtained under the full model but $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ are fitted under the reduced model. The benefit is that the likelihoods under the full and reduced models are comparable, nuisance from refitting FPCA is avoided, and the likelihood ratio is guaranteed to be positive. The marginal likelihood based test statistic (GLR-ML) is defined as

$$T_{n,\mathcal{M}} = \widehat{l}_{n,\mathcal{M}}(H_1) - \widehat{l}_{n,\mathcal{M}}(H_0).$$

Similarly, we can define GLR statistics based on conditional likelihood (GLR-CL)

$$T_{n,\mathcal{C}} = \widehat{l}_{n,\mathcal{C}}(H_1) - \widehat{l}_{n,\mathcal{C}}(H_0),$$

where both likelihoods are as defined in (4.8) using the same σ^2, eigenfunctions and FPC scores estimated under the full model, and the difference is that we plug in full model estimator for $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ in $\widehat{l}_{n,\mathcal{C}}(H_1)$ and reduced model mean estimators in $\widehat{l}_{n,\mathcal{C}}(H_0)$.

For comparison, we also define a test based on working independence which totally ignores the covariance structure among the response variables. In general, WI is a

simple strategy in longitudinal data analysis that results in consistent estimation (Lin and Carroll, 2001) and legitimate test procedures (Tang, Li, and Guan, 2016). In fact, our initial mean estimators in Section 4.3.1, now denote as $\widehat{\boldsymbol{\mu}}_g^{\mathcal{W}}$ and $\widehat{\boldsymbol{\alpha}}^{\mathcal{W}}$, are WI estimators. The WI test statistic (GLR-WI) is defined as

$$T_{n,\mathcal{W}} = \widehat{l}_{n,\mathcal{W}}(H_1) - \widehat{l}_{n,\mathcal{W}}(H_0),$$

where $\widehat{l}_{n,\mathcal{W}}(H_1) \propto -\frac{1}{2}\sum_{g=1}^{G} \|\boldsymbol{Y}_g - \widehat{\boldsymbol{\mu}}_g^{\mathcal{W}} - \mathbf{X}_g\widehat{\boldsymbol{\alpha}}^{\mathcal{W}}\|^2$ and $\widehat{l}_{n,\mathcal{W}}(H_0)$ is defined similarly except to replace the mean estimators by their reduced model counterparts. The WI test is easy to implement since we can skip the FPCA, model selection and the refined estimation procedure in Section 4.3.3. However, we find in our numerical studies that the GLR tests based on $T_{n,\mathcal{M}}$ and $T_{n,\mathcal{C}}$, both of which rely on FPCA, yield higher power than the WI test.

We propose to estimate the null distribution of the proposed test statistics using a permutation strategy for the following reasons. First, the asymptotic distributions of the GLR tests on bivariate alternatives verses univariate null have not been investigated in the literature. Second, as pointed out by many authors (Mammen, 1993; Fan and Jiang, 2007; Tang, Li, and Guan, 2016), the asymptotic distribution of a nonparametric test statistic usually performs poorly under finite samples because of the slow convergence rate in nonparametric settings. Even for the cases where the asymptotic distribution of the test statistic is available, these authors favor resampling methods.

A simple permutation strategy is to break the association between the response $Y_i(t)$ and the lunar day s_i. The test procedure is given as follows.

Step 1: Randomly permute the lunar day s_i to s_i^* such that all seeds measured on the same day in the original data still have the same lunar day in the permuted data set. The permuted data set can be expressed as $\{Y_i(t), \mathbf{X}_i, s_i^*; i = 1\ldots, n\}$.

Step 2: Calculated the GLR test statistic T_n^* based on the permuted data set.

Step 3: Repeat *Step 1* and *Step 2* a large number of times and estimate the p-value by the empirical frequency that T_n^* is greater than T_n.

This procedure is applicable to all three version of T_n proposed above.

4.5 Simulation studies

4.5.1 Results on the estimation procedure

We evaluate the performance of the proposed methodology using simulation studies that mimic the real data. We generate data according to the following model:

$$Y_i(t_j) = \mu(s_i, t_j) + \mathbf{X}_i'\boldsymbol{\alpha} + \sum_{k=1}^{p_1} \xi_{1,g(i),k}\psi_{1,k}(t_j) + \sum_{k=1}^{p_2} \xi_{2,f(i),k}\psi_{2,k}(t_j) + \sum_{k=1}^{p_3} \xi_{3,i,k}\psi_{3,k}(t_j) + \epsilon_i(t_j),$$

where $\mu(s_i, t_j) = \{1 + \cos(2\pi s_i/10)/5\}\{-12(t_j - 1/2)^2 + 3\}$ is the mean function, $\mathbf{X}_i'$ is a vector of camera indicators, $\boldsymbol{\alpha} = (-0.3, -0.2, -0.1, 0.1, 0.2, 0.3)$, $\epsilon_i(t_j) \sim N(0, \sigma^2)$ with $\sigma = 1$, $s_i \in [1, 30]$, $t_j = j/(m_T - 1), j = 0, \ldots, (m_T - 1)$, and $m_T = 31$. The mean function μ, as shown in Figure 4.2 (a), is chosen to mimic the mean function that we obtain from the real data. We simulate data for $n = 1000$ seeds from $G = 100$ genotypes, each genotype consists of two files, and there are five seeds in each file. Following the structure of the real data, we assume that all seeds within a genotype are observed on the same day, and the lunar day of a genotype is set to be random. We consider the following two scenarios and conduct 200 simulations under each scenario.

Scenario I: Let $p_1 = p_2 = p_3 = 2$. The principal component scores $\xi_{l,k}$ are generated independently from $N(0, \omega_{l,k})$, $l = 1, 2, 3$. The eigenvalues are $(\omega_{1,1}, \omega_{1,2}) = (1, 1/4)$ at the genotype level, $(\omega_{2,1}, \omega_{2,2}) = (1/2, 1/4)$ at the file level, and $(\omega_{3,1}, \omega_{3,2}) = (5, 1/2)$ at the seed level. For the eigenfunctions, we set

$$\psi_{1,1}(t) = \sqrt{2}\sin(2\pi t), \quad \psi_{1,2}(t) = \sqrt{2}\cos(2\pi t), \quad \psi_{2,1}(t) = \sqrt{2}\sin(4\pi t),$$
$$\psi_{2,2}(t) = \sqrt{2}\cos(4\pi t), \quad \psi_{3,1}(t) = 1, \quad \psi_{3,2}(t) = \sqrt{12}(t - 1/2).$$

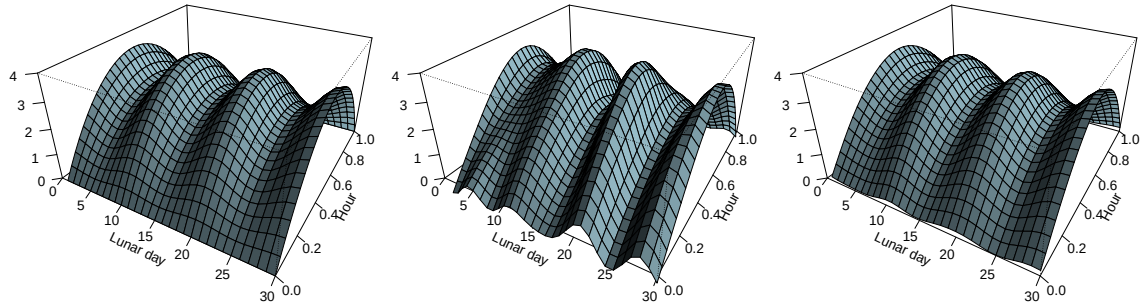
Note that the seed level eigenfunctions are not orthogonal to the genotype and file level eigenfunctions in this scenario.

Scenario Ⅱ: Let $p_1 = 1, p_2 = 1, p_3 = 4$. To illustrate the robustness of our proposed procedure against violation of the normal assumption, we simulate the principal component scores from a skewed Gaussian mixture models. Specifically, for a principal component score ξ with mean zero and variance ω, we generate ξ with probability 1/3 from $N(2\sqrt{\omega/3}, \omega/3)$, and with probability 2/3 from $N(-\sqrt{\omega/3}, \omega/3)$. We set eigenvalues as $\omega_{1,1} = 1/4$ at the genotype level; $\omega_{2,1} = 1/2$ at the file level; and $(\omega_{3,1}, \omega_{3,2}, \omega_{3,3}, \omega_{3,4}) = (2, 1, 1/2, 1/4)$ at the seed level. For the eigenfunctions, we consider the following mutually orthogonal functions:

$$\psi_{1,1}(t) = \sqrt{2}\sin(2\pi t), \quad \psi_{2,1}(t) = \sqrt{2}\cos(2\pi t), \quad \psi_{3,1}(t) = \sqrt{2}\sin(4\pi t),$$
$$\psi_{3,2}(t) = \sqrt{2}\cos(4\pi t), \quad \psi_{3,3}(t) = \sqrt{2}\sin(6\pi t), \quad \psi_{3,4}(t) = \sqrt{2}\cos(6\pi t).$$

We first focus on the estimation results under Scenario I. To estimate $\mu(s,t)$, we use the tensor products between $K_T = 14$ cubic B-splines in t (10 interior knots and 4 boundary knots) and $K_L = 11$ Fourier functions in s as the basis. We also use the tensor product of 14 B-splines for all covariance estimation. All tuning parameters, including λ_μ and ϱ in (4.5) and $\lambda_{\mathcal{G}1}$, $\lambda_{\mathcal{G}2}$ and $\lambda_{\mathcal{G}3}$ for covariance estimation, are chosen by GCV. In Figure 4.2, we compare the true function $\mu(s,t)$ with the proposed spline estimator in a typical run and the mean of our estimator averaged over 200 runs. As we expected, compared with the working independence estimator, the refined estimator of the mean function using our iterative procedure improves the estimation in terms of integrated mean square error (23.12% reduction under Scenario I and 28.28% reduction under Scenario Ⅱ). We also provide box plots of the eigenvalues in Figure 4.3 and a graphical summary for the estimated eigenfunctions in Figure 4.4, where we compare in each panel the 2.5% and 97.5% point-wise percentiles of our eigenfunction estimator with the truth. As we can see, all of these estimators perform quite well, the estimated

eigenvalues are close to the true values, and the true eigenfunctions are always nicely covered by the percentile bands. Figure 4.5 shows that the predicted principal component scores are very close to true principal component scores. We observe in Figure 4.3 that the estimated eigenvalues in the third level have less bias than the first two levels. Similarly, in Figure 4.4, the percentile bands for the third level eigenfunctions are tighter than the first two levels. An explanation of this phenomenon is that there are more repetitions for the third level functional random effects ($n = 1000$) than the first two levels ($G = 100$ and $F = 200$). Estimation results for Scenario Ⅱ are similar to those in Scenario I and hence relegated to the Supplementary Material.



(a) (b) (c)

Figure 4.2 Estimation of the mean surface based on 200 simulations under Scenario I. Panel (a) is the true mean surface; Panel (b) is the estimated mean surface based on one simulation; Panel (c) is the estimated mean surface based on the average of 200 simulations.

4.5.2 Model selection results

To evaluate the finite sample performance of our model selection procedure based on conditional AIC, we compare our method with the widely-used PVE method (Di et al., 2009; Li, et al., 2015). The threshold percentage in the PVE method is usually chosen subjectively, e.g. Di et al. (2009) use 90% and Li, et al. (2015) use 85%. To make a fair comparison, we consider both thresholds in our simulations for the PVE method. For each simulation, we choose (p_1, p_2, p_3) by searching the minimum of the conditional

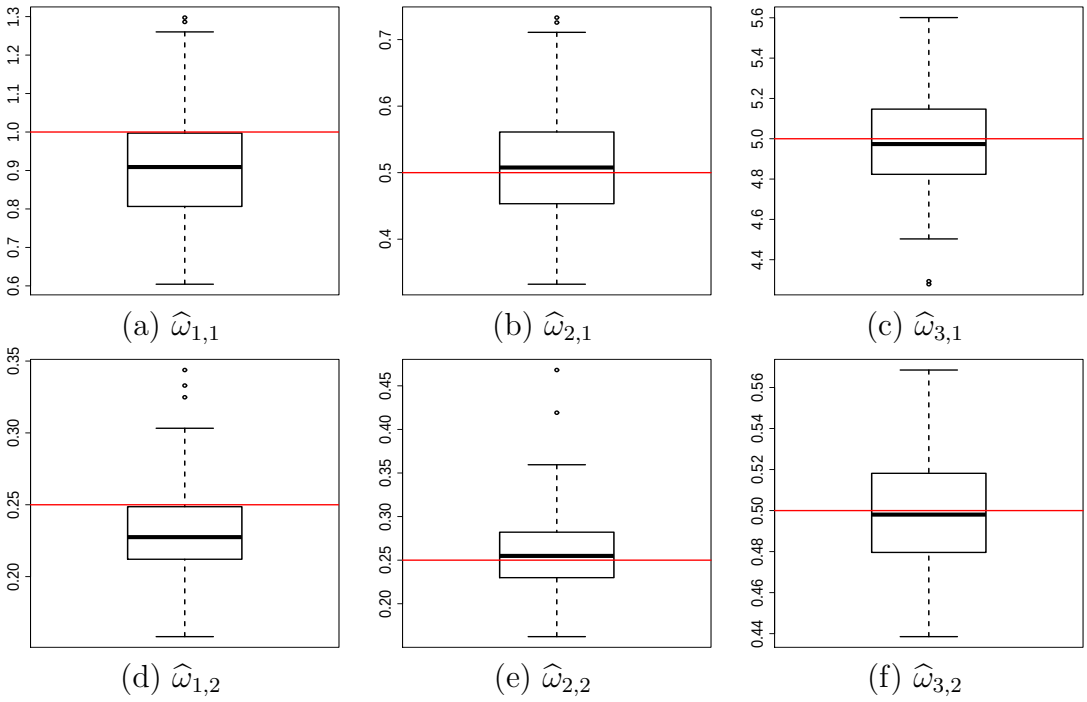


Figure 4.3 Boxplots of the estimated eigenvalues based on the 200 simulations under Scenario I. The solid lines are the true eigenvalues.

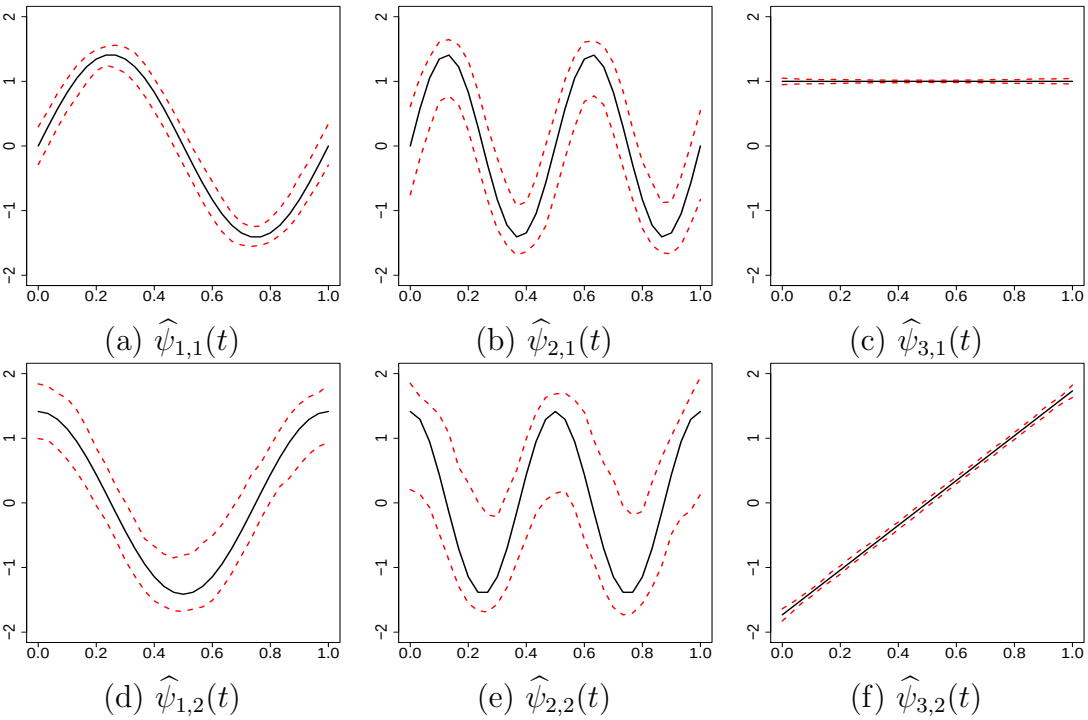
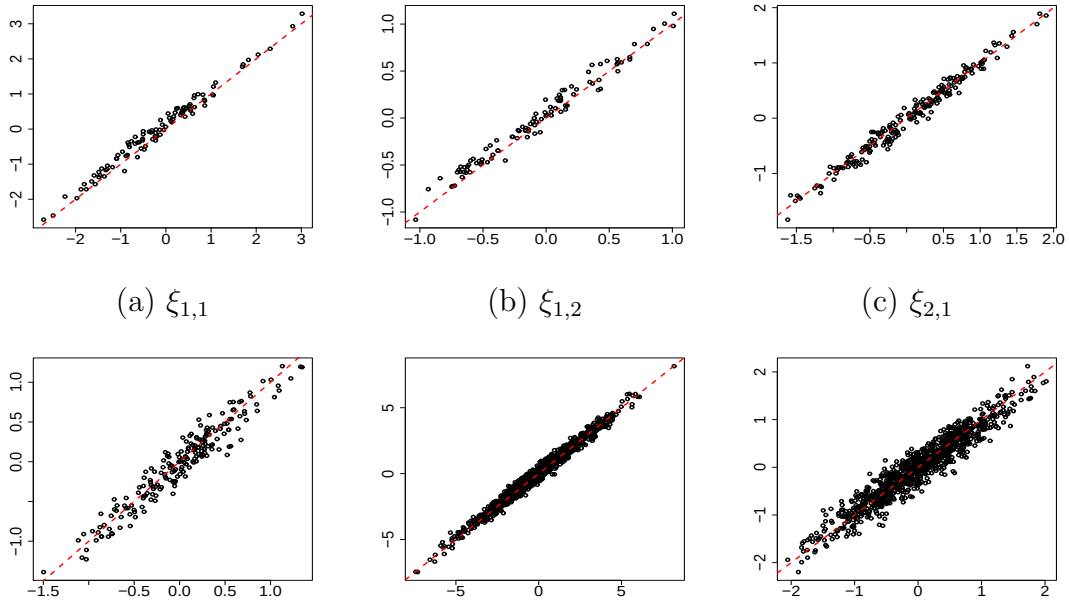


Figure 4.4 The true eigenfunctions and their corresponding 95% confidence bands based on point-wise 2.5% and 97.5% quantiles of the 200 simulations under Scenario I.



(a) $\xi_{1,1}$ (b) $\xi_{1,2}$ (c) $\xi_{2,1}$

(d) $\xi_{2,2}$ (e) $\xi_{3,1}$ (f) $\xi_{3,2}$

Figure 4.5 Predicted principal component scores against true principal component scores for the first simulated data set under Scenario I. The dashed lines are 45 degree reference lines.

AIC using a simple grid search method. The model selection results are summarized in Table 4.1, where we present the empirical distribution of the selected number of principal components in each level of the hierarchy by each method. The mode frequency of the each estimator is marked in bold. Under Scenario I, our AIC picks the correct number of principal component 100%, 97.5% and 100% of the time for the three hierarchies respectively; the PVE method choose the wrong model for level 2 and level 3 very often. Overall, our proposed method have 97.5% of chance choosing the correct number of principal components in all levels, while PVE method fails miserably in this category. Under Scenario II with the correct model $(p_1, p_2, p_3) = (1, 1, 4)$, the contrast between our method and the PVE method is even more striking: our method selects the correct number of components in all levels of the hierarchy 100% of the time, while PVE method misses the true model most of the time in almost all levels regardless which threshold value is in use.

Table 4.1 Empirical distributions for the number of selected principal components $\widehat{p}$ for the three levels of hierarchy using various methods.

Criteria	Level	$\widehat{p}=1$	$\widehat{p}=2$	$\widehat{p}=3$	$\widehat{p}=4$	All levels
Scenario I: $p_1=p_2=p_3=2$						
AIC	1	0.000	**1.000**	0.000	0.000	**0.975**
	2	0.000	**0.975**	0.025	0.000	
	3	0.000	**1.000**	0.000	0.000	
PVE 85%	1	0.005	**0.970**	0.025	0.000	0.000
	2	0.000	**0.815**	0.185	0.000	
	3	**1.000**	0.000	0.000	0.000	
PVE 90%	1	0.000	**0.930**	0.070	0.000	0.095
	2	0.000	**0.685**	0.315	0.000	
	3	**0.840**	0.160	0.000	0.000	
Scenario Ⅱ: $(p_1,p_2,p_3)=(1,1,4)$						
AIC	1	**1.000**	0.000	0.000	0.000	**1.000**
	2	**1.000**	0.000	0.000	0.000	
	3	0.000	0.000	0.000	**1.000**	
PVE 85%	1	0.425	**0.575**	0.000	0.000	0.000
	2	**0.640**	0.360	0.000	0.000	
	3	0.000	0.000	**1.000**	0.000	
PVE 90%	1	0.190	**0.740**	0.070	0.000	0.000
	2	0.345	**0.645**	0.010	0.000	
	3	0.000	0.000	**1.000**	0.000	

Note: The empirical distributions above are based on 200 simulation runs. The last column contains the frequency of choosing the correct numbers of principal components in all levels.

It is also worth noting that the data under Scenario Ⅱ are non-Gaussian. Even though our conditional AIC is motivated from a conditional Gaussian likelihood and the principal component scores are estimated using BLUP under Gaussian assumption, the procedure performs remarkably well under Scenario Ⅱ. These results show that both the BLUP for the PC scores and the AIC are robust against mild violation of Gaussian assumptions.

4.5.3 Hypothesis test results

We now demonstrate the performance the proposed GLR tests on the hypotheses in (4.10). We first investigate whether the proposed permutation procedure retains the

nominal size of the test and compare the powers of the three version of GLR tests. Since both GLR-ML and GLR-CL depend on specification of the numbers of principal components in FPCA, we also investigate the sensitivity of our GLR test statistics to these choices.

We adopt the simulation setting in Scenario I described in Section 4.5.1 but set the mean function to be $\mu(s_i, t_j) = \{1 + \delta \cos(2\pi s_i/10)/5\}\{-12(t_j - 1/2)^2 + 3\}$ for some constant δ. It is easy to see that the null hypothesis $\mu(s,t) \equiv \mu(t)$ is true when $\delta = 0$, and larger value of δ indicates further deviation from the null. We simulate 200 data sets for each $\delta \in \{0, 0.5, 1, 1.5, 2\}$ so that data are generated under both the null and alternative hypotheses.

We first assume the numbers of principal components are correctly specified, and perform level $\alpha = 0.05$ tests on the null hypothesis (4.10) for each simulated data set, where the decision is made using the permutation procedure described in Section 4.4.2. The empirical powers of the three GLR tests, as functions of δ, are shown in Figure 4.6 (a). As we can see, the permutation method estimate the null distributions remarkably well and all three GLR tests hold their nominal size. By comparing the three power curves, it seems that the GLR-ML is most powerful, followed by GLR-CL and then GLR-WI. The gap between the three test is the largest when $\delta = 1$. To confirm this observation, we use McNemar's test to test if the powers of two tests are the same at $\delta = 1$ and the p-values are $3 * 10^{-8}$ for GLR-ML vs. GLR-CL and 0.02 for GLR-CL vs. GLR-WI. This finding supports our intuition that modeling the covariance structure in functional data would increase the power of tests.

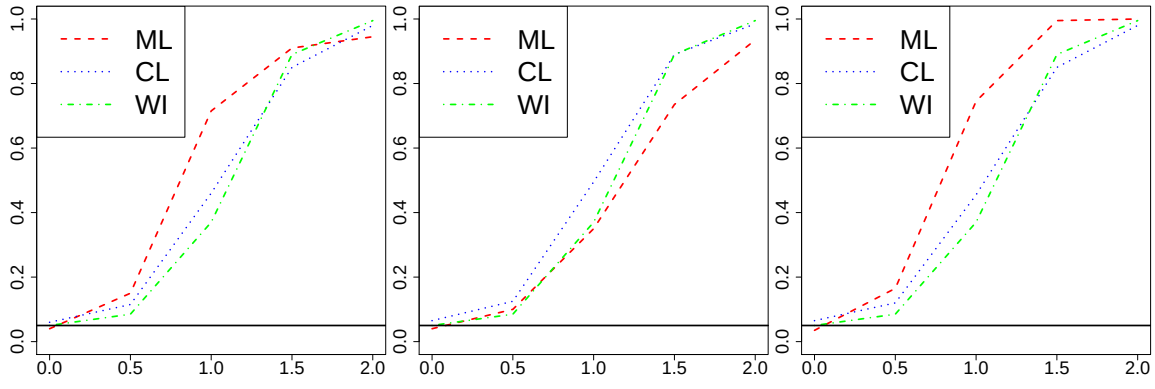
Next, we investigate the sensitivity of the test statistics to the selected number of principal components. Figure 4.6 (b) shows the powers of the three GLR tests when the numbers of principal components are underestimated as $(p_1, p_2, p_3) = (1, 1, 1)$; and Figure 4.6 (c) shows the powers when the numbers of principal components are overestimated as $(p_1, p_2, p_3) = (3, 3, 3)$. It is interesting to see that the permutation method maintains

the nominal sizes for all tests even under model mis-specification. Since GLR-WI does not depend on FPCA, its power curve is identical across the three panels of Figure 4.6. When the number of principal components are under estimated, GLR-ML looses a lot of power and can fall under GLR-WI; GLR-CL is relatively robust and still more powerful than GLR-WI.

When the numbers of principal components are overestimated, panel (c) of Figure 4.6 is almost identical to panel (a), showing both GLR-ML and GLR-CL are robust against over estimating the numbers of components. One simple explanation is that, when more principal components are selected which means additional eigenvalues and eigenfunctions are included, the variances of those additional eigenvalues are usually relatively small which might not have a significant effect on the power of the tests. However, if less principal components are selected, then the missing eigenvalues are the true ones, so the result of missing them could be disastrous, especially for the marginal likelihood-based test. Although selecting more principal components may not have a big impact in terms of testing, it makes the model more complicated than necessary, and the additional principle components which are artificial noises lose interpretability.

4.6 Data analysis

We now apply our proposed methodology to our motivating data, the RIS data. The implementation details of our methodology are described in Section 4.3, Section 4.4, and Section 4.5. Figure 4.7 (a) shows the empirical mean surface by averaging bending rates over each grid of a lunar day and observation time. The two missing stripes correspond to the 7th and the 21st lunar day when no experiments were conducted. On each observed lunar day, the empirical mean curve looks like a parabola with the highest bending rate around the half time of the 1.5 hours. However, for each observation time, the empirical mean curve is neither complete nor smooth, so it is difficult to judge whether the bending

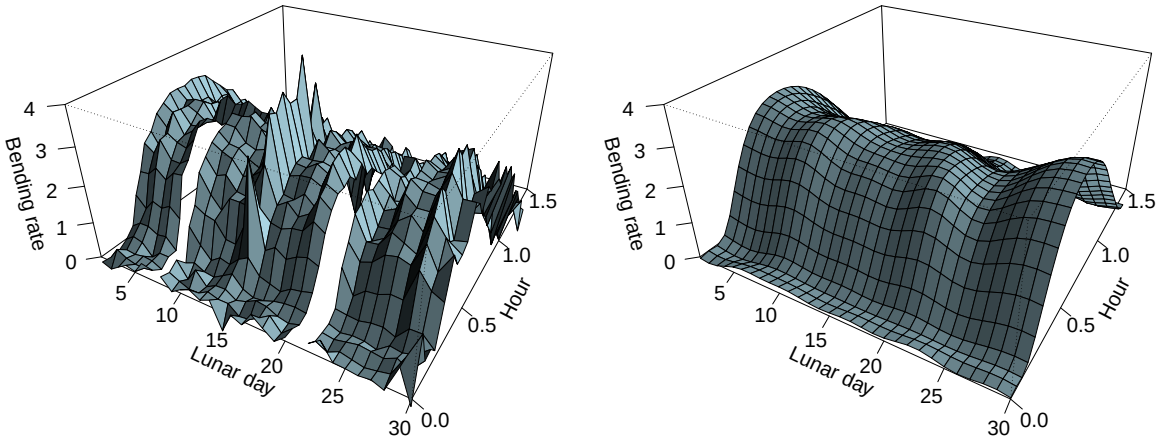


(a) (b) (c)

Figure 4.6 Performance and sensitivity of the GLR tests. X-axis: δ; y-axis: empirical power based on 200 simulations; Solid line: the reference line at 0.05; Dashed line: marginal likelihood (ML) based GLR test; Dotted line: conditional likelihood (CL) based GLR test; Dot-dashed line: working independent (WI) GLR test. Panel (a): model is correctly specified with $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3) = (2, 2, 2)$; Panel (b): the numbers of principal components are underestimated as $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3) = (1, 1, 1)$; Panel (c): the numbers of principal components are overestimated as $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3) = (3, 3, 3)$.

rate depends on the lunar day variable or not. Figure 4.7 (b) shows the estimated mean surface, which is smooth and probably imply that the mean function is a bivariate function which depends on both the lunar day variable and observation time. We will formally investigate this hypothesis later using our proposed GLR tests. According to Figure 4.7 (b), the highest peak is around the new moon and the other two high peaks are near the first quarter and the third quarter. We treated the 7th camera as the reference. The estimate of the camera effect is $\widehat{\boldsymbol{\alpha}} = (-0.092, -0.041, -0.017, -0.028, -0.068, -0.016)'$, so the effect of camera setups is relatively small.

We applied our proposed conditional AIC to choose the numbers of principal components (p_1, p_2, p_3). Our proposed method chooses the model $(1, 1, 4)$, while the PVE method chooses the model $(2, 2, 3)$ with 85% variation explained and $(3, 3, 3)$ with 90% variation explained. The results of both methods indicate that the information of all three levels is contained in very low dimensions. Table 4.2 displays the numbers of principal components chosen for the three levels using our method and the corresponding



(a) (b)

Figure 4.7 The mean surface estimates based on the RIS data. (a) is the empirical mean surface; (b) is the estimated mean surface using our proposed method.

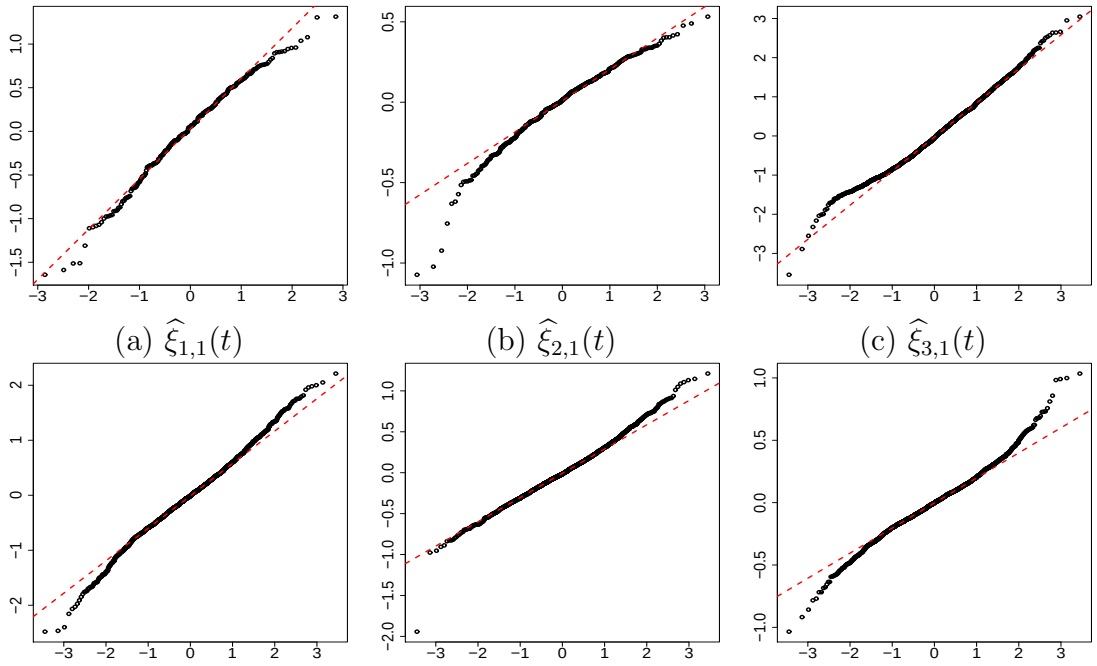
Table 4.2 Estimated eigenvalues for the three levels for the RIS data. "percent var" means the percentage of variance explained by a component, and "cum percent var" means the cumulative percentage of variance explained

	Level 1	Level 2	Level 3			
Component	1	1	1	2	3	4
Eigenvalue	0.324	0.101	0.699	0.499	0.132	0.079
Percent var	56.515	50.854	51.226	29.996	9.702	5.804
Cum percent var	56.515	50.854	51.226	81.222	90.924	96.728

estimated eigenvalues. In terms of estimated eigenvalues, most information is contained in the third (seed) level and least information is contained in the second (file) level. To be more specific, the genotype, file and seed level variability account for 18.6%, 5.8%, and 75.6% of variability for the bending rate process. In this regard, the model chosen by our method should be more reasonable because it chooses more principal components in the third level but less in the first two levels compared with the PVE method which chooses the numbers of principal components independently in all levels using a same percentage. The estimate of the white noise measurement error variance is $\widehat{\sigma}^2 = 1.029$ which is a fair amount of noise.

Figure 4.8 shows Q-Q plots for the estimated principal component scores for the three levels based on our selected model. The distributions of the principal component scores are slightly heavy-tailed or skewed but close to normal. As shown in Section 4.5, our method are robust when the distributions of the principal component scores slightly deviate from normality. Corresponding to the estimated principal component scores, Figure 4.9 displays estimated eigenfunctions for all levels. Figure 4.9 (a) shows the estimated eigenfunction for the genotype. The function is generally negative with the lowest value at about 1 hour. It indicates that seeds with negative scores on the genotype component will tend to have higher bending rate than the population average and this tendency becomes strongest at about 1 hour after the observation. This U-shaped estimated eigenfunction coincides with our common sense because the genetic effects are expected to be small at the beginning and the end of the selected observation period and to be large around the middle of the period. Figure 4.9 (b) shows the estimated eigenfunction for the file which is similar to a sinusoidal function. For files with positive scores, the file effects is negative during the first half period and positive during the second half period. Figure 4.9 (c) shows the estimated eigenfunction corresponding to the first principal component score of the seed. It is very similar to the estimated eigenfunction for the file level except a sign change. Figure 4.9 (d) - (f) show the remaining estimated eigenfunctions for the seed level which are also close to sinusoidal functions.

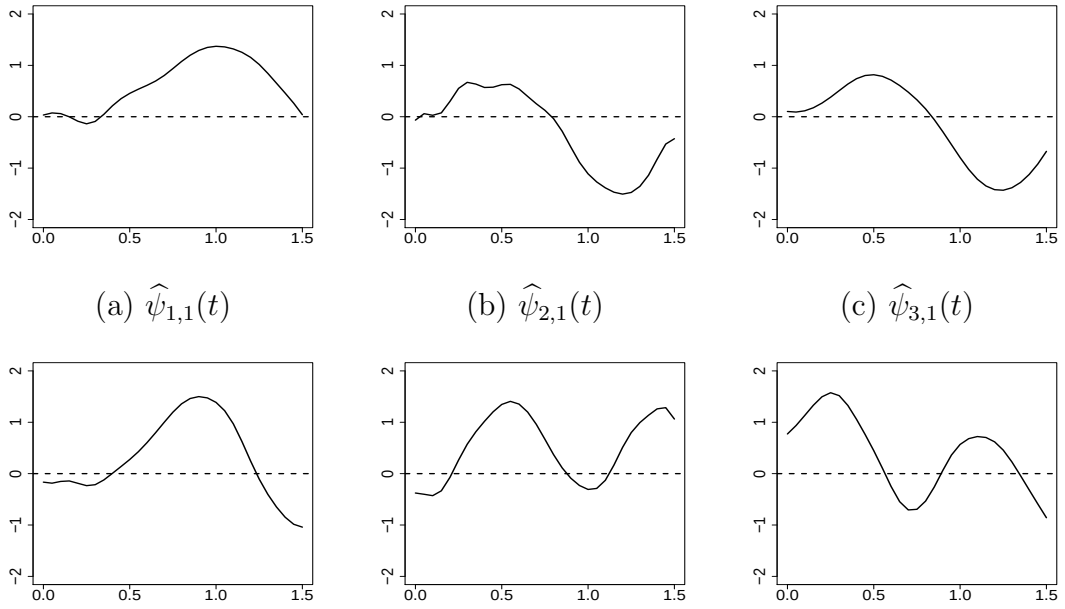
Part of the motivation of our research is to answer an important scientific question about whether the moon phase effect exists. Intuitively, the mean surface in Figure 4.7 (b) supports the existence of the effect, but a more formal conclusion is needed using hypotheses testing. To answer this question, we used the proposed GLR tests described in Section 4.4.2. To approximate the null distributions of the GLR statistics, we used the proposed permutation strategy with 1000 repetitions. The working independence based test and the conditional likelihood based test yield a p-value of 0.025 and 0.043 which indicate that the moon phase effect is significant. The marginal likelihood based test



(a) $\widehat{\xi}_{1,1}(t)$ (b) $\widehat{\xi}_{2,1}(t)$ (c) $\widehat{\xi}_{3,1}(t)$

(d) $\widehat{\xi}_{3,2}(t)$ (e) $\widehat{\xi}_{3,3}(t)$ (f) $\widehat{\xi}_{3,4}(t)$

Figure 4.8 Q-Q plots for the predicted principal component scores.



(a) $\widehat{\psi}_{1,1}(t)$ (b) $\widehat{\psi}_{2,1}(t)$ (c) $\widehat{\psi}_{3,1}(t)$

(d) $\widehat{\psi}_{3,2}(t)$ (e) $\widehat{\psi}_{3,3}(t)$ (f) $\widehat{\psi}_{3,4}(t)$

Figure 4.9 Estimation of the eigenfunctions based on the data of the RIS.

gives a p-value of 0.204 which shows some discrepancy. We believe the moon phase effect exists because of the following reasons. Firstly, the working independent test, which is the most robust because it does not depend on model selection, shows that the effect is significant. Secondly, the conditional likelihood based test, which is more robust against model misspecification compared with the marginal likelihood based test, also indicates significant moon phase effect. Finally, although the marginal likelihood based test is the most powerful under the correct model, additional investigation in the simulation studies shows that under the alternative hypothesis it is not rare that for some samples both the working independence based test and the conditional likelihood based test reject the null hypothesis but the marginal likelihood based test cannot.

4.7 Discussion

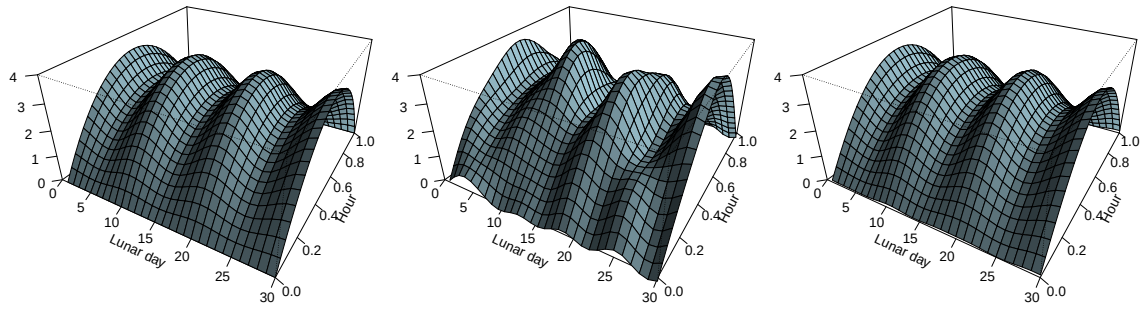
Motivated by the RIS data, we have proposed methodology for estimation, model selection, and testing in the presence of 3-level nested function data. Our methodology is proposed under a very general framework in the sense that it can be easily generalized to be applied to more complicated multi-level functional data. For example, it is straightforward to modify our methodology and apply it to 3-level correlated function data (Li, et al., 2015) or 4-level hierarchical function data. The RIS data have a natural 3-level, genotype-file-seed, nested hierarchical structure. However, since there is only one principal component selected for the file level and it only accounts for 5.8% variability of the total variability of the 3 levels, it is an interesting problem to investigate in future about whether the file level is necessary. Our data analysis shows that the camera effect is relatively small, it is of interest to know whether the camera effect exists. Wald test (Li, et al., 2015) could be developed to test the parametric part in our model. To define the bending rate process, we take consecutive differences which may introduce some noise. An alternative way to model the bending rate process is to use the original data

directly and model the functional derivative (Liu and Müller, 2009) of the response using multilevel FPCA. However, this is outside of the scope of this paper and is left as an open problem.

To estimate the null distributions of the GLR statistics, we propose a simple permutation strategy based on the hierarchical structure of the data. This strategy works well in terms of holding the nominal size of the test and retaining good power. In the simulation studies, as we expected, the likelihood-based tests based on FPCA have better power than the working independence based test. Sensitivity analysis for the proposed two likelihood-based GLR tests also has been conducted. The methods and principals we propose are applicable to broader settings in functional data analysis.

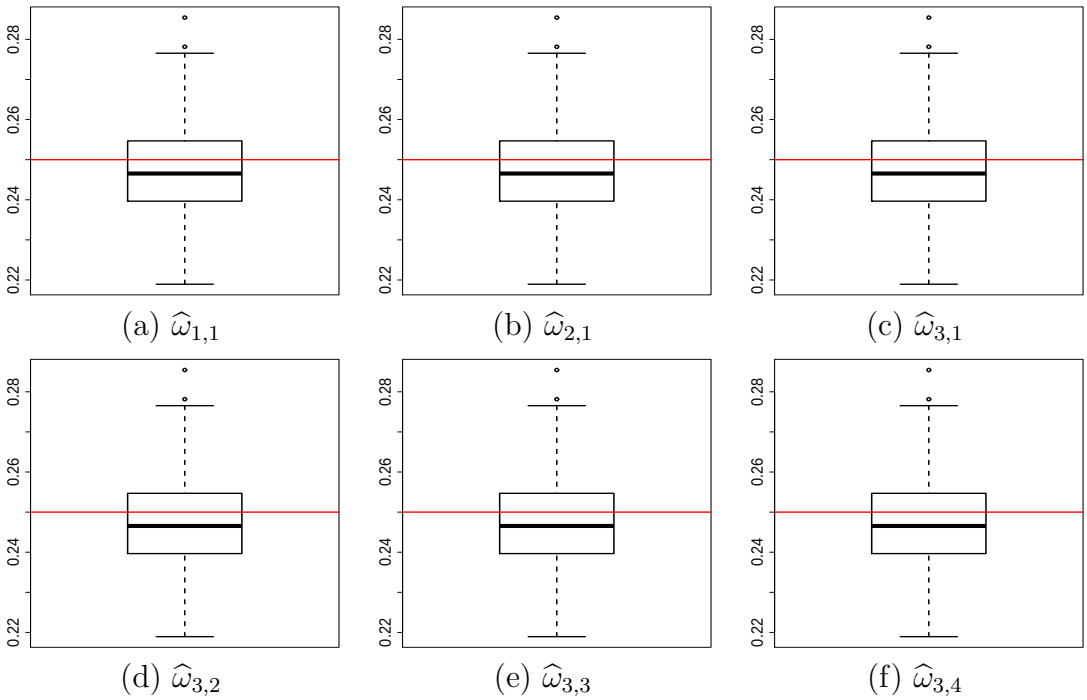
4.8 Supplementary materials

In the supplementary materials,we provide additional simulation results for Scenario Ⅱ described in Section 4.5 and additional plots for data analysis.



(a) (b) (c)

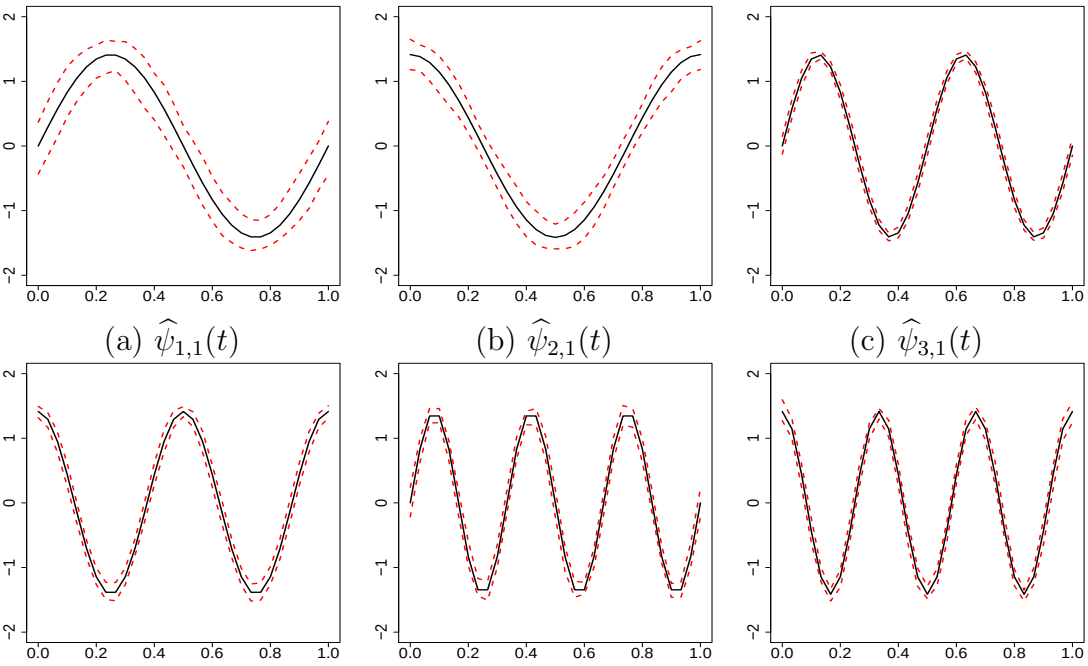
Figure 4.10 Estimation of the mean surface based on 200 simulations under Scenario Ⅱ. Panel (a) is the true mean surface; Panel (b) is the estimated mean surface based on one simulation; Panel (c) is the estimated mean surface based on the average of 200 simulations.



(a) $\widehat{\omega}_{1,1}$ (b) $\widehat{\omega}_{2,1}$ (c) $\widehat{\omega}_{3,1}$

(d) $\widehat{\omega}_{3,2}$ (e) $\widehat{\omega}_{3,3}$ (f) $\widehat{\omega}_{3,4}$

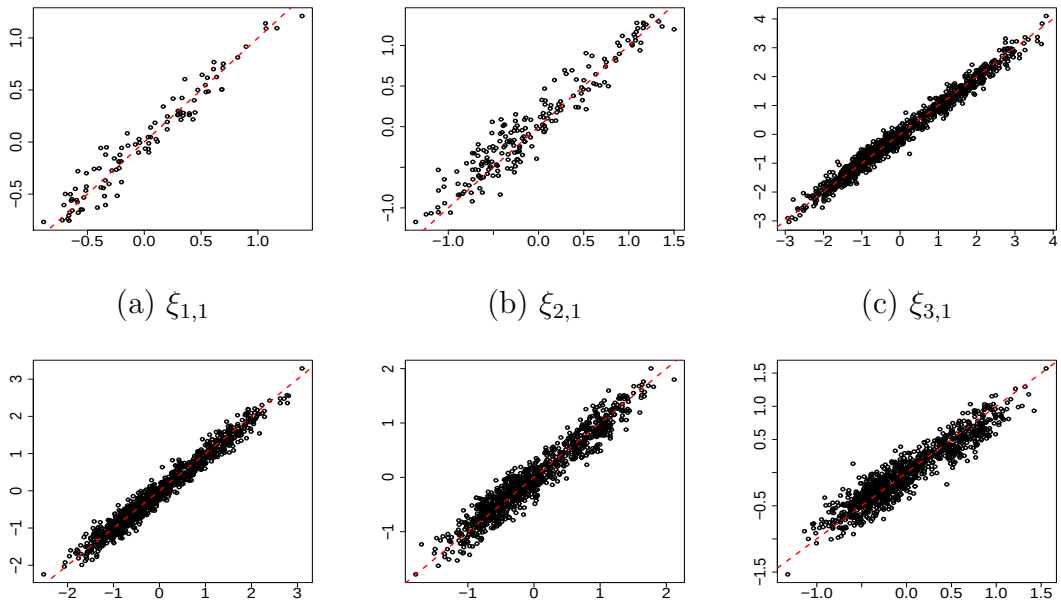
Figure 4.11 Boxplots of the estimated eigenvalues based on the 200 simulations under Scenario Ⅱ. The solid lines are the true eigenvalues.



(a) $\widehat{\psi}_{1,1}(t)$ (b) $\widehat{\psi}_{2,1}(t)$ (c) $\widehat{\psi}_{3,1}(t)$

(d) $\widehat{\psi}_{3,2}(t)$ (e) $\widehat{\psi}_{3,3}(t)$ (f) $\widehat{\psi}_{3,4}(t)$

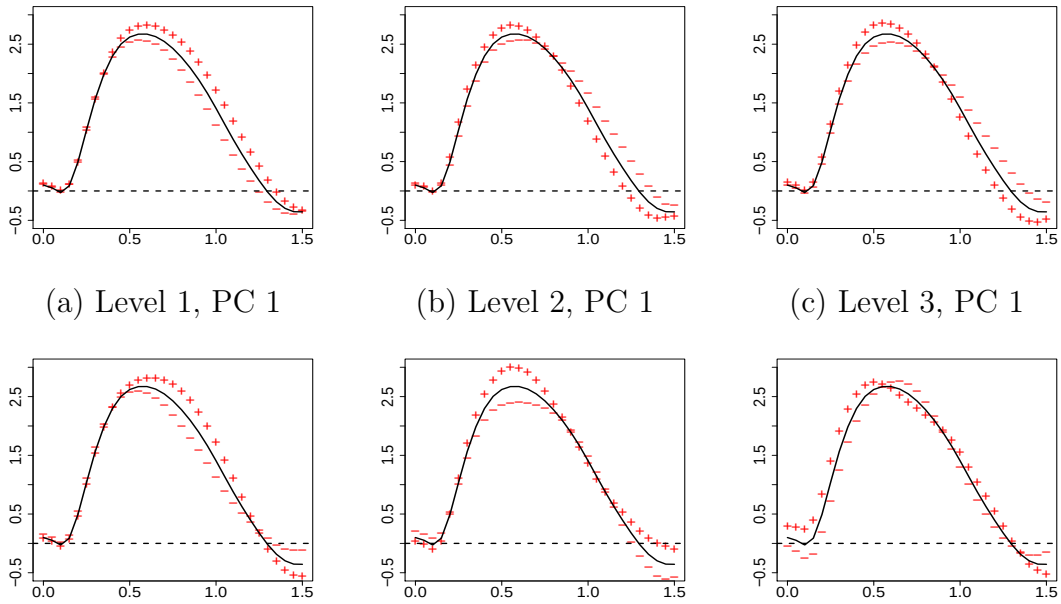
Figure 4.12 The true eigenfunctions and their corresponding 95% confidence bands based on point-wise 2.5% and 97.5% quantiles of the 200 simulations under Scenario Ⅱ.



(a) $\xi_{1,1}$ (b) $\xi_{2,1}$ (c) $\xi_{3,1}$

(d) $\xi_{3,2}$ (e) $\xi_{3,3}$ (f) $\xi_{3,4}$

Figure 4.13 Predicted principal component scores against true principal component scores for the first simulated data set under Scenario Ⅱ. The dashed lines are 45 degree reference lines.



(a) Level 1, PC 1 (b) Level 2, PC 1 (c) Level 3, PC 1

(d) Level 3, PC 2 (e) Level 3, PC 3 (f) Level 3, PC 4

Figure 4.14 The mean bending rate curves and the effects of adding, denoted by "+", and subtracting, denoted by "−", a suitable multiple of each PC curve.

CHAPTER 5. SUMMARY AND DISCUSSION

In this thesis, we present two different topics in measurement error problems (Chapter 2 and Chapter 3) and one topic in multilevel functional data analysis (Chapter 4). The main contributions and some discussions about our proposed methods are summarized as follows.

In Chapter 2, we propose locally efficient estimators for proportional hazards models with measurement error. Although there are lots of methods proposed on measurement error problems for the proportional hazards models (see Carroll et al. (2006)), none of them yields semiparametric efficient estimators. We propose novel estimators which prove to be locally efficient in the sense that the estimators are semiparametrically efficient if the distribution of the error-prone covariates is specified correctly, and are still consistent and asymptotically normal if the distribution is misspecified. We generalize Tsiatis and Ma (2004) which focuses only on functional measurement error models to a more complex and general setup and prove some asymptotic properties of our estimators. We also propose a sandwich formula for the variance estimation of our estimators and it works well in our simulation studies. Our numerical studies show that our method vastly outperforms competing methods. The methodology proposed in this chapter can be applied to other statistical areas, such as modeling interval-censored data and joint modeling of longitudinal and time-to-event data.

In Chapter 3, we propose semiparametric estimators for general regression problems when error-prone surrogates of the true predictors are collected in the primary data set while accurate measurements of the predictors are available only in a small validation

data set. There have been some likelihood or score equation based semiparametric methods proposed based on this topic, but none of them yield consistent estimators for the regression coefficients under a general regression setup. In contrast, we show that our proposed estimators based on expected estimating equations (Wang and Pepe, 2000) are consistent and robust, and can be applied to a general regression model. Our data analysis which motivates our study indicates the importance of keeping a proper BMI to reduce the risk of acquiring hypertension. The proposed method is based on kernel smoothing which might suffer from the curse of dimensionality, so we suggest to use some variable selection or feature screening methods to circumvent the curse of dimensionality when the predictors are high-dimensional.

In Chapter 4, we propose methodology on the estimation, model selection and nonparametric testing on the mean function for a nested hierarchical functional data model. Although there has been some recent research on hierarchical or multi-level functional principal component analysis, the following contributions make our work unique. Firstly, our proposed iterative algorithm for the mean function estimation is computationally more efficient than the computationally intensive EM algorithm which is used in Li, et al. (2015). Secondly, our proposed method to choose the number of principal components based on a conditional AIC is completely data-driven and vastly outperforms the existing ad hoc methods based on the findings in the simulation studies. Finally, to test the existence of moon phase effect, we propose novel nonparametric tests based on GLR. Similar types of tests have not been investigated before. We also compare the power of the three proposed tests and conduct sensitivity analysis to study the robustness of the three tests under model misspecification. Significant moon phase effect has been discovered by applying our proposed tests to the data on the root image study. Our proposed methodology can be generalized to analyze more complex 3-D image data, such as fMRI data.

BIBLIOGRAPHY

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective (2nd ed.)*. London: Chapman and Hall CRC Press.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error model. *Journal of the Royal Statistical Society, Series B*, *53*, 652–663.

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *Annals of Statistics*, *33*, 414–436.

Chatterjee, N. and Chen, Y. (2007). A semiparametric pseudo-score method for analysis of two-phase studies with continuous phase-I covariates. *Lifetime Data Analysis*, *13*, 607–622.

Cheng, Y.-J. and Crainiceanu, C. M. (2009). Cox models with smooth functional effects of covariates measured with error. *Journal of the American Statistical Association*, *104*, 1144-1154.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, *34*, 187–220.

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *Annals of Statistics*, *36*, 665–685.

Di, C., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel Functional Principal Component Analysis. *Annals of Applied Statistics*, *3*, 458–488.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the bootstrap*. London: Chapman and Hall CRC Press.

Fan, J. and Gijbels, I. (1996). *Local Polynomial modeling and its applications*. London: Chapman and Hall CRC Press.

Fan, J. and Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests (with discussion). *Test*, *16*, 409–478.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with Discussion). *Journal of the Royal Statistical Society, Series B*, *70*, 849–911.

Fan, J. and Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics*, *29*, 153–193.

Fuller, W. A. (1987). *Measurement error models*. New York: John Wiley.

González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *Test*, *22*, 361–411.

Hall, P., Müller, H. G., and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, *34*, 1493–1517.

Hammer, S. M., Katezstein, D. A., Hughes, M. D., Gundaker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S, and Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine*, *335*, 1081–1090.

Hastie, T. and Tibshirani, R. (1990). *Generalised additive models*. London: Chapman and Hall CRC Press.

Horn, R. and Johnson, C. R. (1985). *Matrix analysis*. New York: Cambridge University Press.

Hu, P., Tsiatis, A. A., and Davidian, M. (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics*, *54*, 1407–1419.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics*, *24*, 540–568.

Huang, J., Horowitz, J. L., and Wei F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, *38*, 2282–2313.

Huang, Y. and Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: a nonparametric correction approach. *Journal of the American Statistical Association*, *95*, 1209–1219.

James, G., Hastie, T., and Sugar, C. (2000). Principal component models for sparse functional data. *Biometrika*, *87*, 587–602.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, *98*, 119–132.

Kim, J. K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. New York: Chapman and Hall CRC Press.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith S. M., Subar, A. F., Tooze, J. A., Carroll, R. J., and Freedman, L. S. (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, *65*, 1003–1010.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association*, *90*, 78–94.

Lee, L. F. and Sepanski, J. H. (1995). Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association*, *90*, 130–140.

Li, Y., Guolo, A., Hoffman, F. O., and Carroll, R. J. (2007). Shared uncertainty in measurement error models, with application to Nevada test site fallout data. *Biometrics*, *63*, 1226–1236.

Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*, *38*, 3321-3351.

Li, H., Keadle, S. K., Staudenmayer J., Assaad, H., Huang, J. Z., and Carroll, R. J. (2007). Methods to assess an exercise intervention trial based on 3-level functional data. *Biostatistics*, *16*, 754–771.

Li, Y., and Lin, X. H. (2003). Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. *Journal of the American Statistical Association*, *98*, 191–203.

Li, Y., Wang, N., and Carroll, R. J. (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association*, *108*, 1284–1294.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, *96*, 1045–1056.

Liu, B. and Müller, H. G. (2009). Estimating Derivatives for Samples of Sparsely Observed Functions With Application to Online Auction Dynamics. *Journal of the American Statistical Association*, *104*, 704–717.

Ma, Y. and Carroll, R. J. (2006). Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association*, *101*, 1465–1474.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, *21*, 255–285.

Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, *48*, 829–838.

Noh, B., Bandyopadhyay, A., Peer, W. A., Spalding, E. P., and Murphy, A. S. (2003). Enhanced gravi-and phototropism in plant mdr mutants mislocalizing the auxin efflux protein PIN1. *Nature*, *423*, 999–1002.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables*. San Diego: Academic Press.

Pepe, M. S. and Fleming, T. R. (1991). A general nonparametric method for dealing with errors in missing or surrogate covariate data. *Journal of the American Statistical Association*, *86*, 108–113.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, *69*, 331–342.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis (2nd ed.)*. New York: Springer-Verlag.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, *82*, 299–314.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Schumaker, L. L. (1981). *Spline functions*. New York: Wiley.

Sepanski, J. H. and Lee, L. F. (1995). Semiparametric estimation of nonlinear error-in-variables models with validation study. *Journal of Nonparametric Statistics*, *4*, 365–394.

Serban, N. and Jiang, H. (2012). Multilevel functional clustering analysis. *Biometrics*, *68*, 805-814.

Shen, X. (1997). On methods of sieves and penalization. *Annals of Statistics*, *25*, 2555–2591.

Song, X., Davidian, M., and Tsiatis, A.A. (2002). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics*, *3*, 511–528.

Song, X. and Huang, Y. (2005). On corrected score approach for proportional hazards model with covariate measurement error. *Biometrics*, *61*, 702–714.

Stone, C. J., Hansen, M., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics*, *25*, 1371–1470.

Stute, W., Xue, L., and Zhu, L. (2007). Empirical likelihood inference in nonlinear errors-in-covariables models with validation data. *Journal of the American Statistical Association*, *102*, 332–346.

Su, Y. R. and Wang, J. L. (2012). Modeling left-truncated and right-censored survival data with longitudinal covariates. *Annals of Statistics*, *40*, 1465–1488.

Tang, J., Li, Y., and Guan, Y. (2016). Generalized quasi-likelihood ratio tests for semiparametric analysis of covariance models in longitudinal data. *Journal of the American Statistical Association*, In press.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. New York: Springer-Verlag.

Tsiatis, A. A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, *88*, 447–458.

Tsiatis, A. A. and Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, *91*, 835–848.

van der Vaart, A. W. (1998). *Asymptotic statistics*. New York: Cambridge University Press.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wang, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *Annals of Statistics*, *39*, 389–417.

Wang, C. Y., Hsu, L., Feng, Z. D., and Prentice, R. L. (1997). Regression calibration in failure time regression. *Biometrics*, *91*, 131–145.

Wang, C. Y. and Pepe, M. S. (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society, Series B*, *62*, 509–524.

Wang, Q. and Rao, J. N. K. (2002). Empirical likelihood-based inference in linear error-in-covariables models with validation data. *Biometrika*, *89*, 345–357.

Wang, C. Y. and Wang, S. (1997). Semiparametric methods in logistic regression with measurement error. *Statistica Sinica*, *7*, 1103–1120.

Wang, Q. and Yu, K. (2007). Likelihood-based kernel estimation in semiparametric errors-in-covariables models with validation data. *Journal of Multivariate Analysis*, *98*, 455–480.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B*, *62*, 413–428.

Xu, Y., Li, Y., and Song, X. (2016). Locally efficient semiparametric estimators for proportional hazards models with measurement error. *Scandinavian Journal of Statistics*, *43*, 558–572.

Yao, F. and Lee, T. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society, Series B*, *68*, 3–25.

Yao, F., Müller, H. G., and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the AmericanStatistical Association*, *100*, 577–590.

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, *95*, 601–619.

Zhou, L., Huang, J., Martinez, J. G., Maity, A., Baladandayuthapani, V., and Carroll, R. J. (2010). Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the American Statistical Association*, *105*, 390–400.

Zhou, S., Shen, X., and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Annals of Statistics*, *26*, 1760–1782.

Zhu, Z., Fung, W. K., and He, X. (2008). On the asymptotics of marginal regression splines with longitudinal data. *Biometrika*, *95*, 907–917.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, *67*, 301–320.